

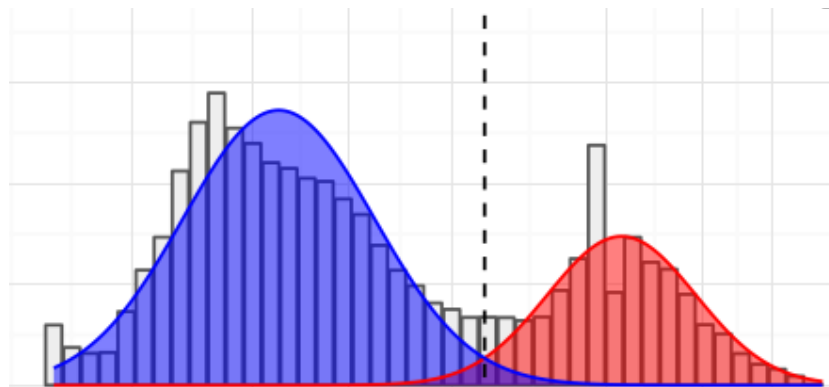
Temporal regularities in activity sessions

The MEGA-STUDY

Aaron Halfaker

ahalfaker@wikimedia.org

User:EpochFail / User:Halfak_(WMF)



Aaron Halfaker

Research Scientist, Wikimedia Foundation

Think big. Measure what you can.



About me

Hi. I'm Aaron Halfaker. I'm a computer scientist. I've been a Wikipedian since 2008. See [my publications](#). In 2011, I started working with the Wikimedia Foundation as a researcher. See my volunteer account, [User:EpochFail](#), for my non-staff work.

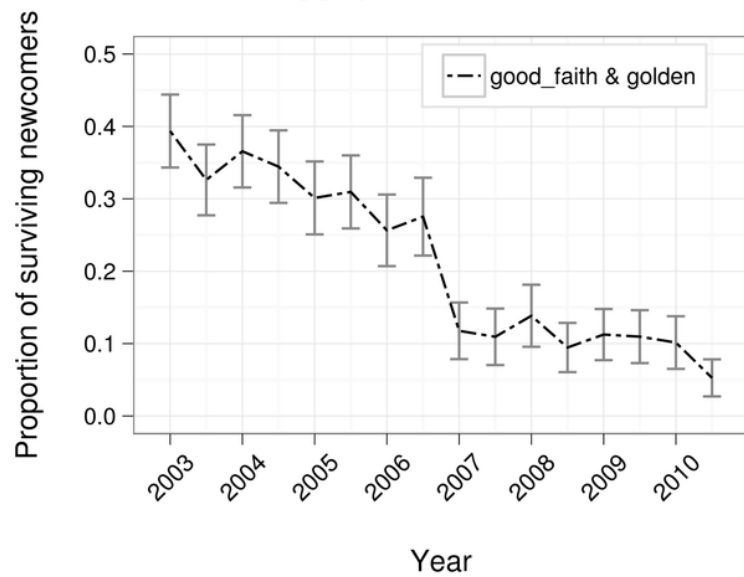
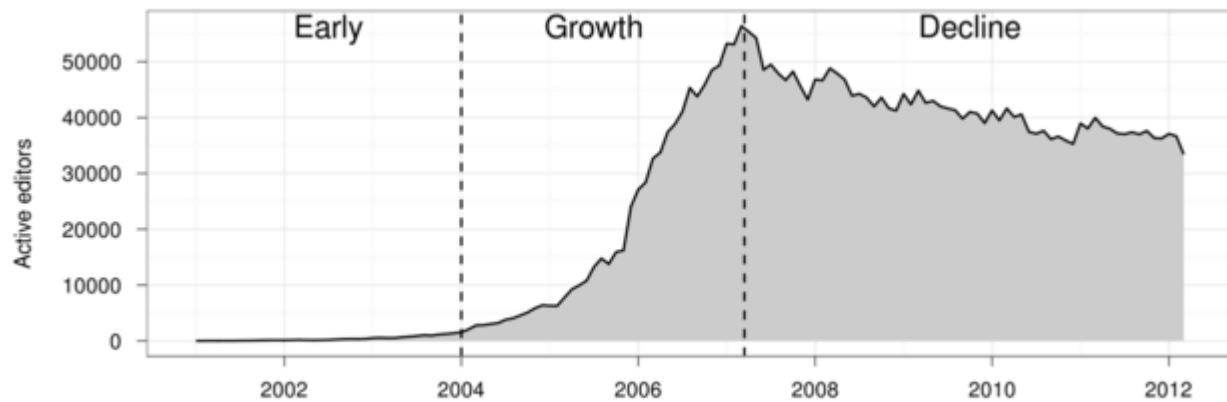
My work

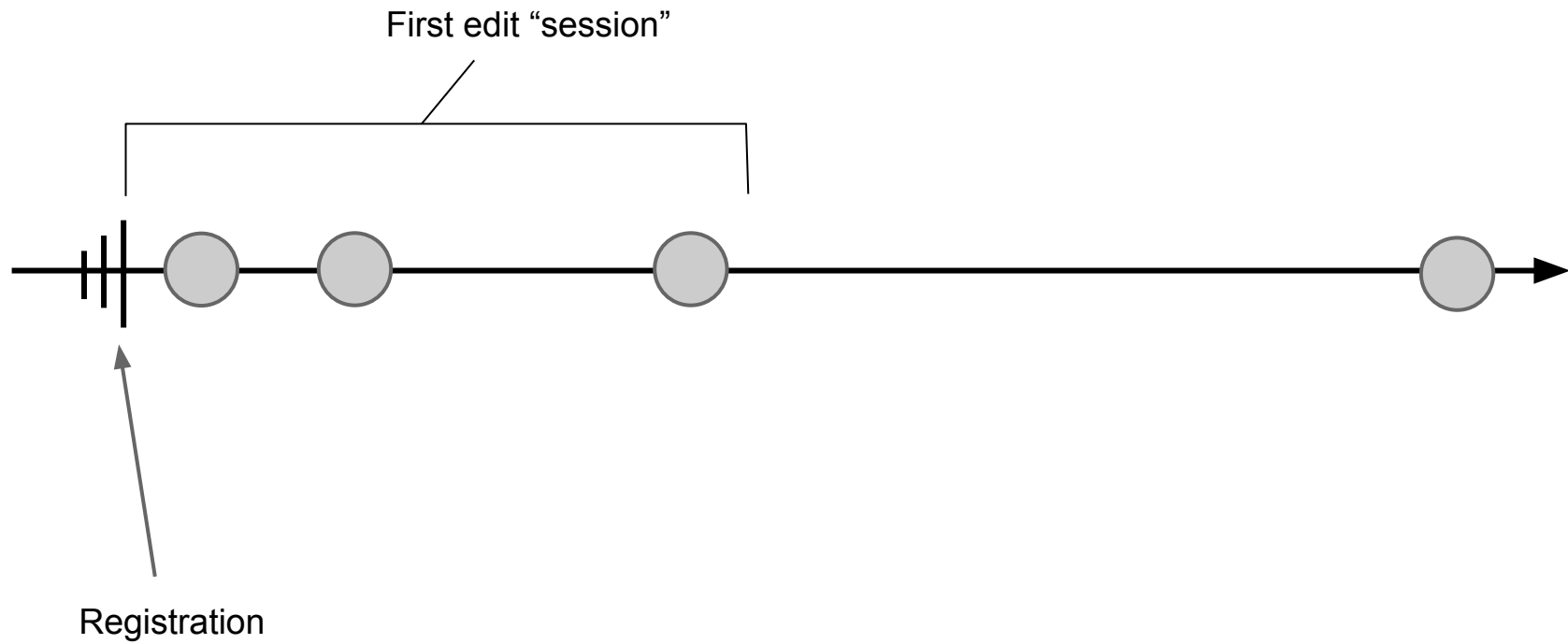
I use the scientific method to explore the structure and function of Wikipedia. My goal is to identify problems/opportunities and to use this knowledge to improve/extend the software. See [The Rise and Decline](#) for an example of my research and [Snuggle](#) for an example of a tool for Wikipedians that I'm actively developing.

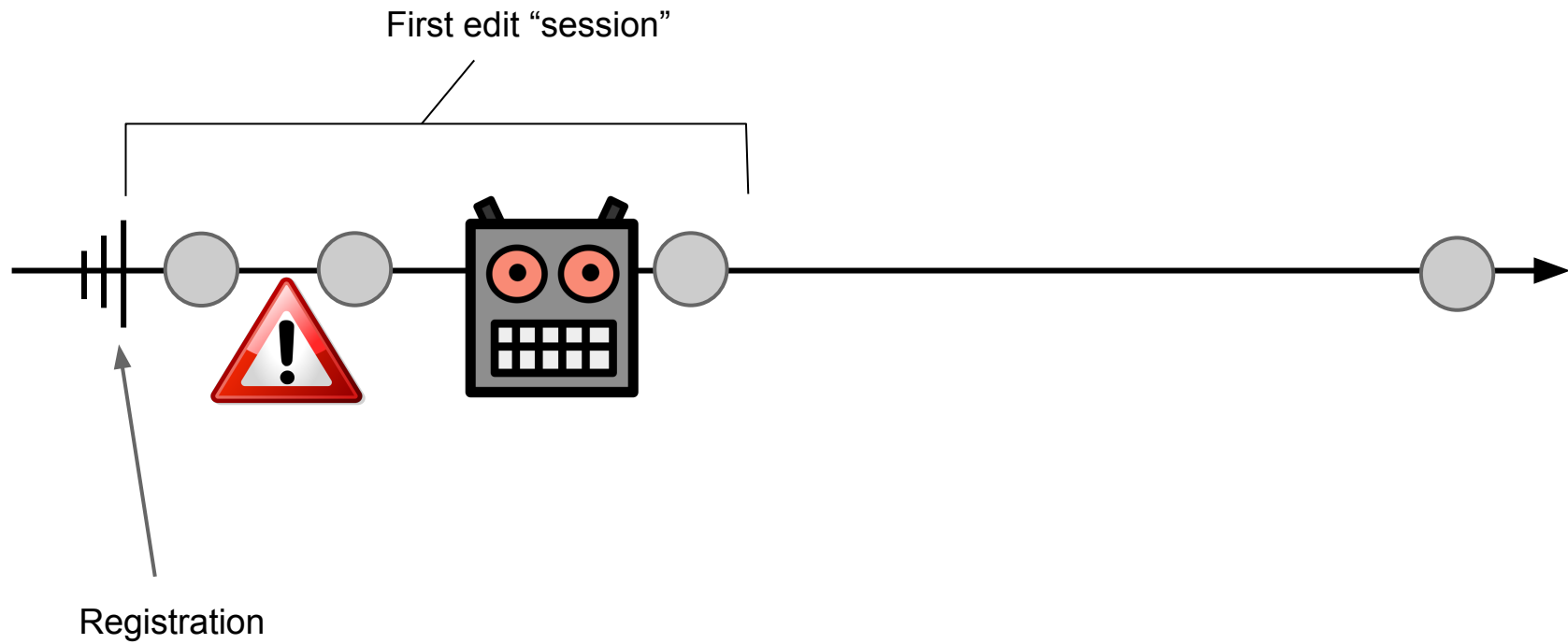
Contact me

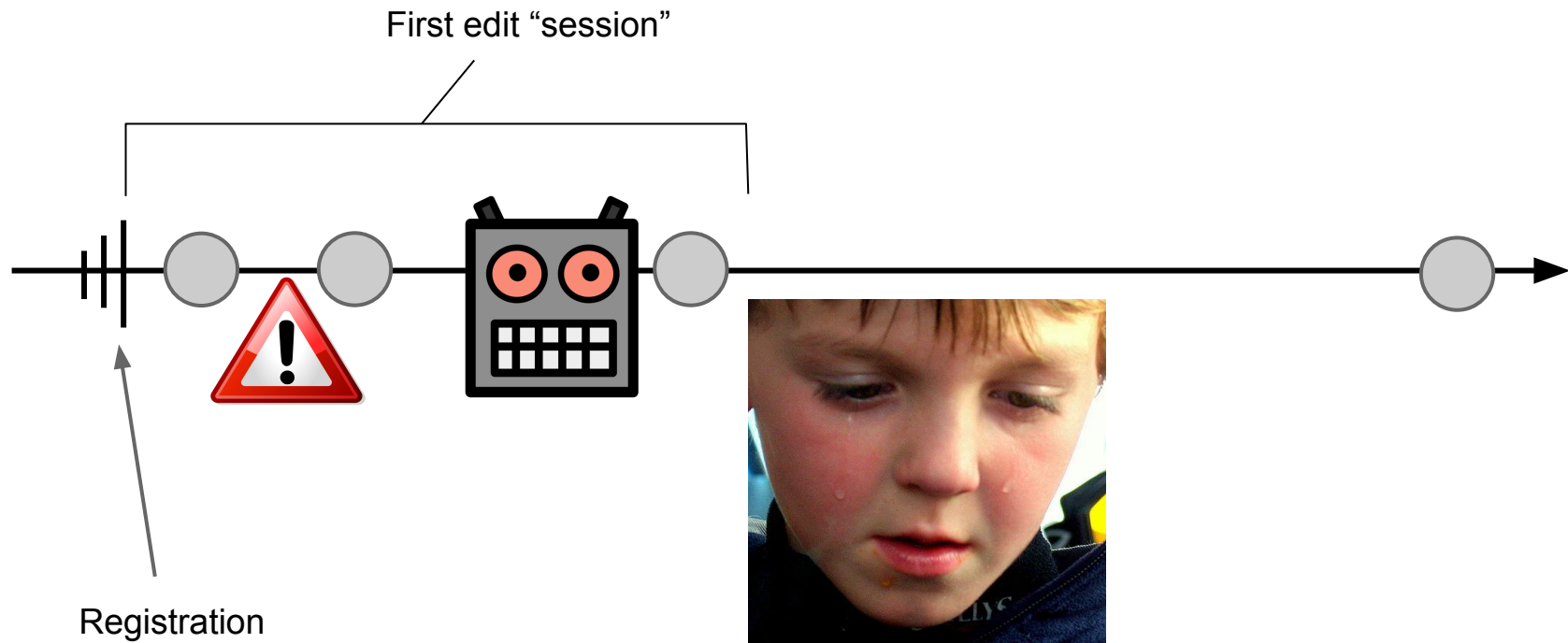
- E-mail: ahalfaker@wikimedia.org
- Website: <http://halfaker.info> [↗](#)
- Twitter: <http://twitter.com/halfak> [↗](#)
- On IRC, I go by "halfak".

Part 1: STORY TIEM!

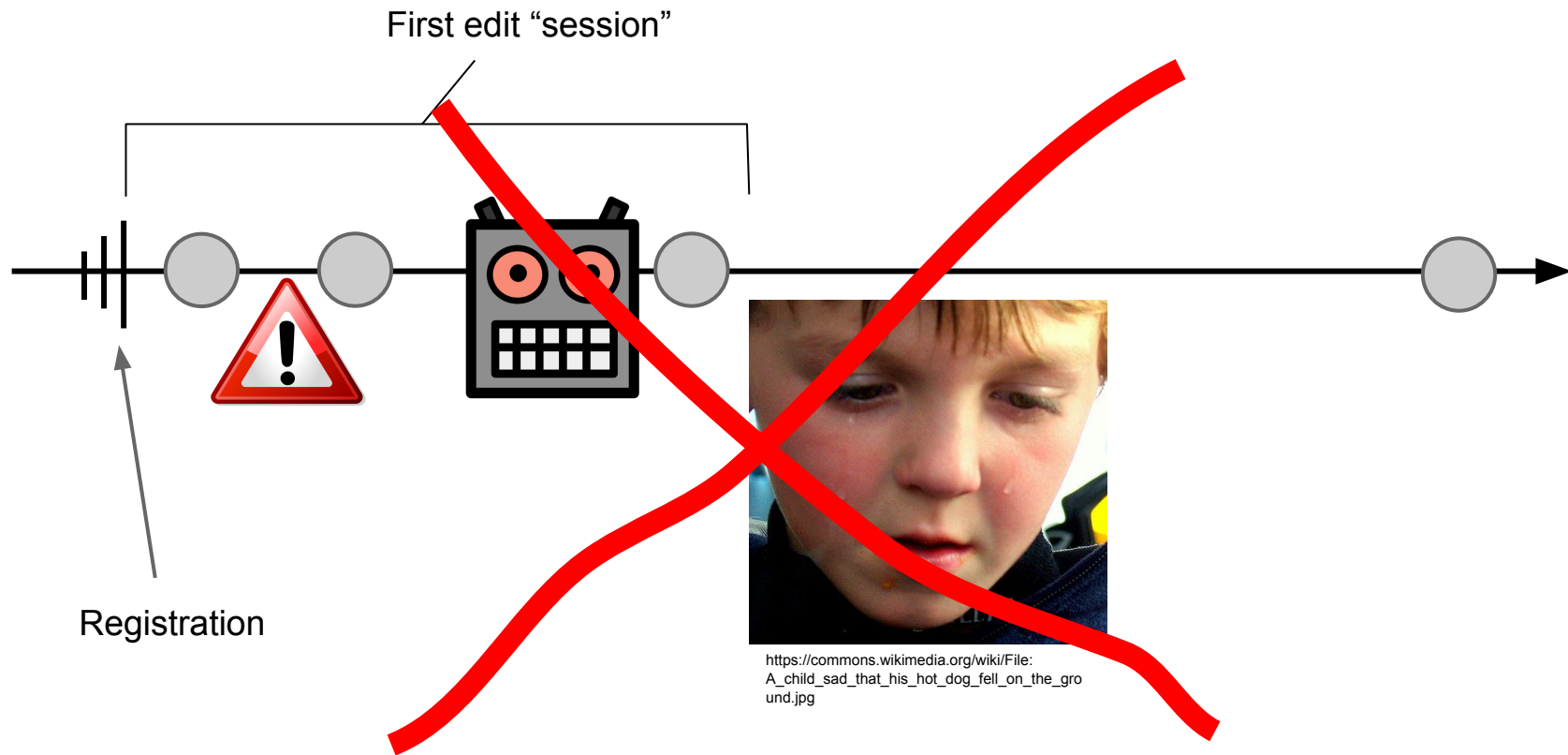








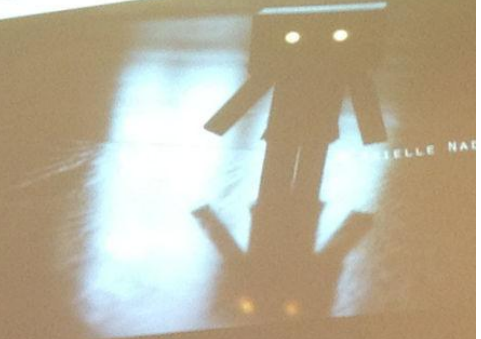
https://commons.wikimedia.org/wiki/File:A_child_sad_that_his_hot_dog_fell_on_the_ground.jpg



...but bots can be a bit scary



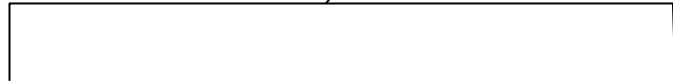
...but bots can be a bit



R. Stuart Geiger

[[User:Staeiou]]

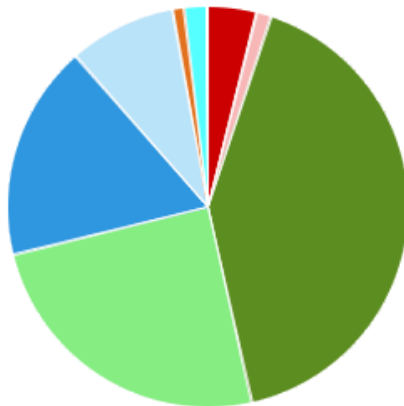
“session”



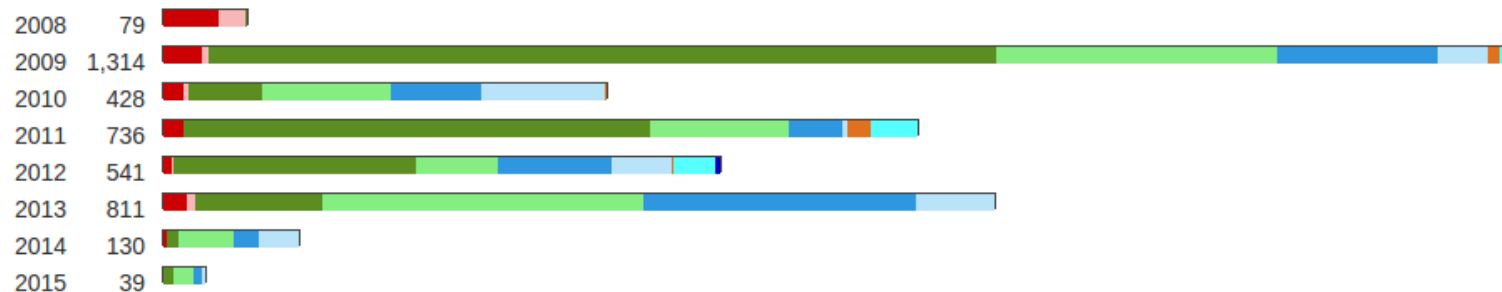
Time spent editing?

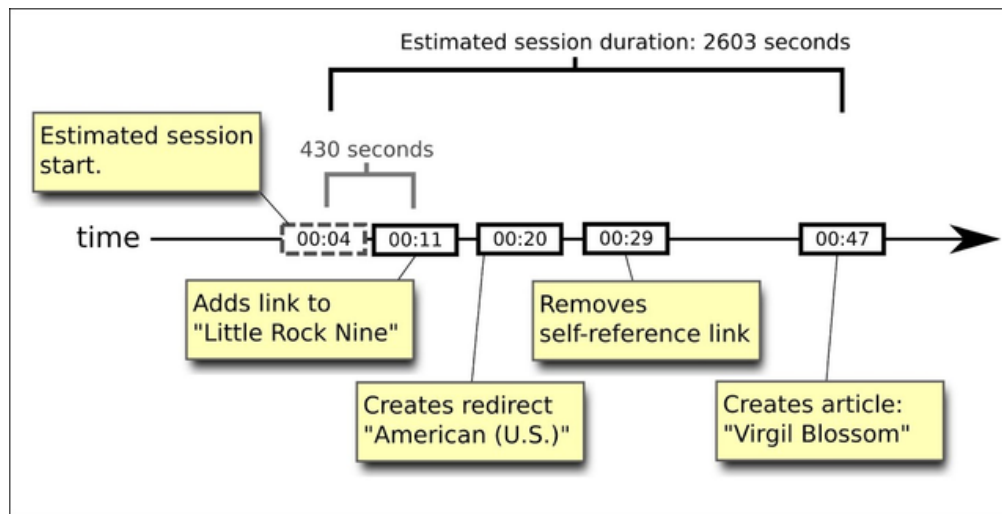
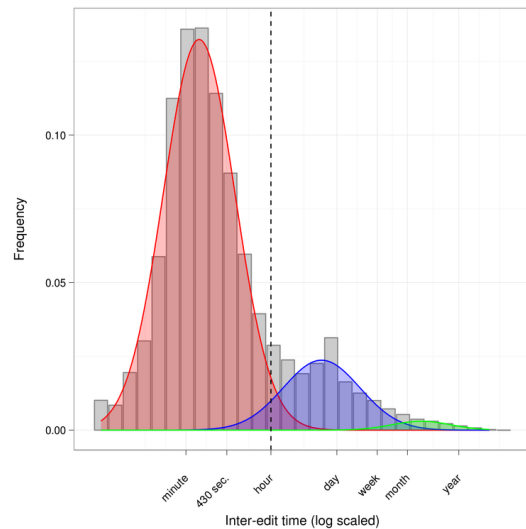
Namespace Totals [hide]

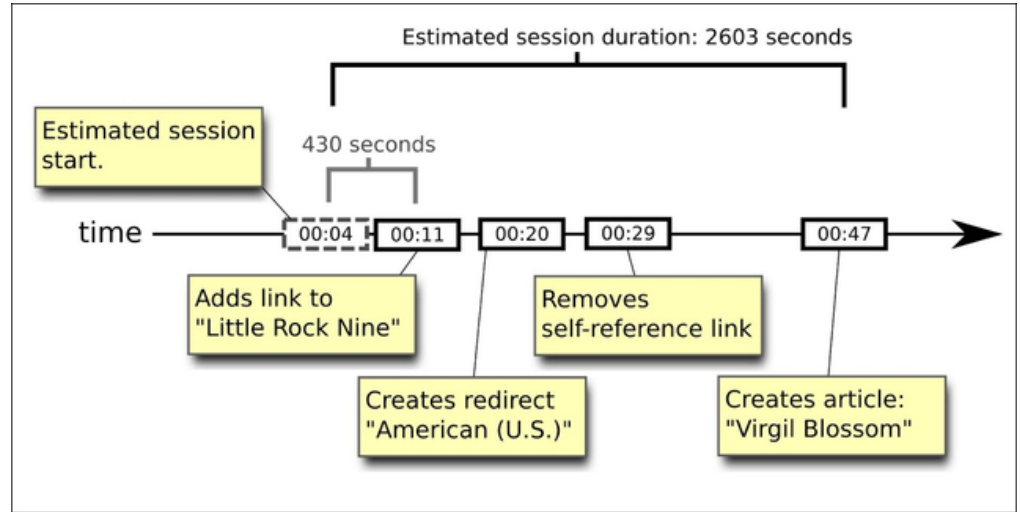
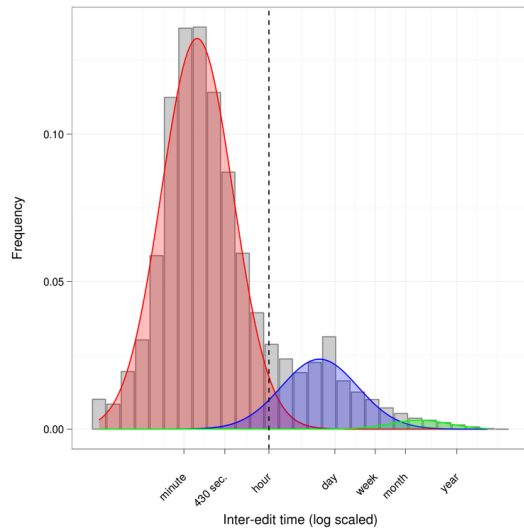
■ Article	161	3.9%
■ Talk	43	1.1%
■ User	1,684	41.3%
■ User talk	1,003	24.6%
■ Wikipedia	707	17.3%
■ Wikipedia talk	347	8.5%
■ File	35	0.9%
■ Template	94	2.3%
■ Template talk	4	0.1%



Year counts [hide]







Channel tunnel

170 million

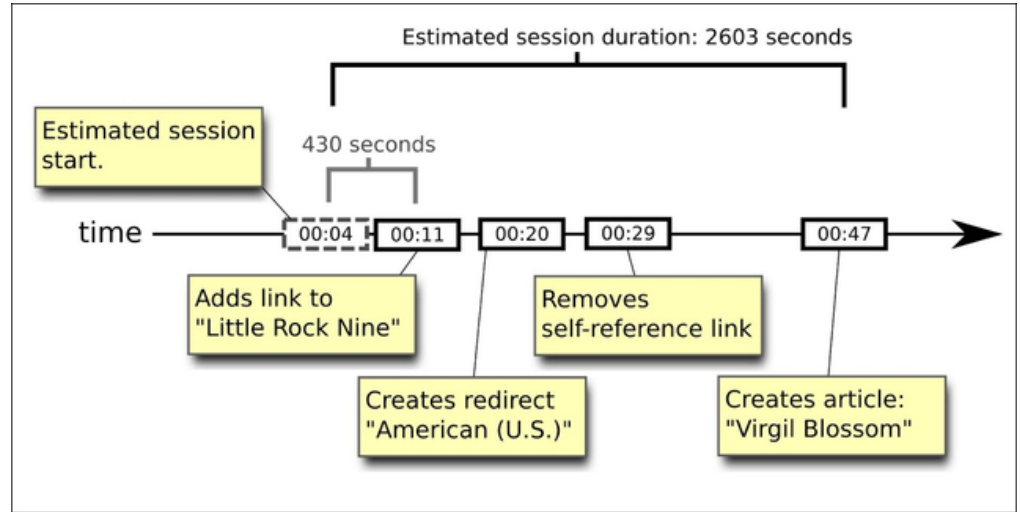
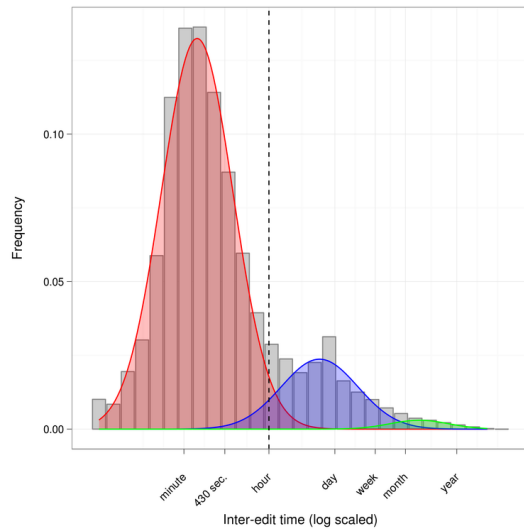
Wikipedia

100 million

Encyclopedia Britannica v1

12 thousand

Geiger, R. S., & Halfaker, A. (2013, February). Using edit sessions to measure participation in Wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 861-870). ACM.



Channel tunnel

170 million

Wikipedia

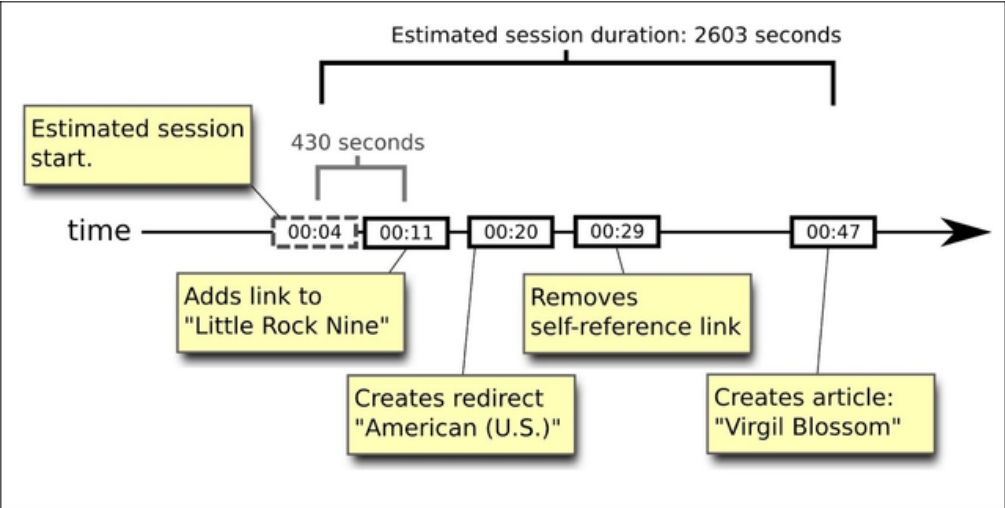
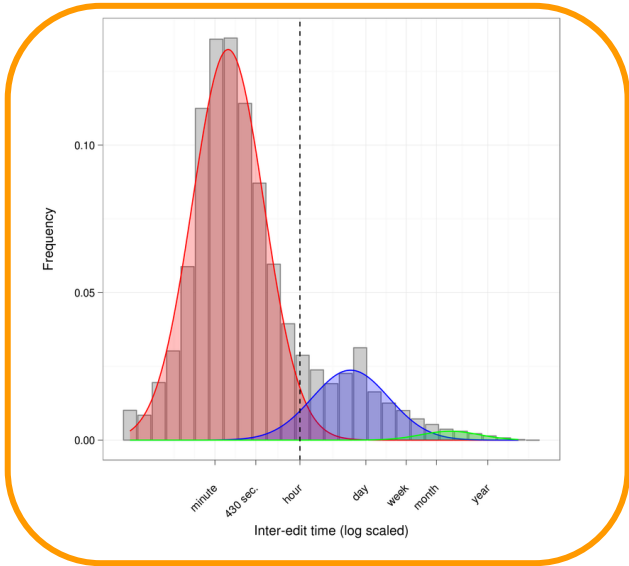
140 million

* based on 12M/year trend

Encyclopedia Britannica v1

12 thousand

Geiger, R. S., & Halfaker, A. (2013, February). Using edit sessions to measure participation in Wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 861-870). ACM.



Channel tunnel

170 million

English Wikipedia

140 million

* based on 12M/year trend

Encyclopedia Britannica

12 thousand

Geiger, R. S., & Halfaker, A. (2013, February). Using edit sessions to measure participation in Wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 861-870). ACM.

Part 2: Methodologies



https://commons.wikimedia.org/wiki/File:See_No_Evil_Hear_No_Evil_Speak_No_Evil.jpg CC-BY-SA 3.0

that 30 minutes produced artefacts around long sessions, and could find no clear evidence of a global session inactivity threshold², while Jones & Klinkner [12] found the 25.5 minute threshold performed “no better than random” in the context of identifying search tasks. Other thresholds have been proposed, but Montgomery and Faloutsos [14] concluded that the actual threshold chosen made little difference to how accurately sessions were identified.

Navigation-oriented heuristics involve inferring browsing patterns based on the HTTP referrers and URLs associated with each request by a user. When a user begins navigating (without a referrer), they have started a session; when a trail can no longer be traced to a previous request based on the referrers and URLs of subsequent requests, the session has ended. This approach was pioneered by Cooley et al in 2002 [5]. While it demonstrated utility in identifying “tasks”, and has been extended by Najjarbashi-Noghani et al. [15], it shows poor performance on sites with framesets due to implicit assumptions about web architecture [3]. Further, the sheer complexity of this strategy and its developmental focus on task over session make it unsuitable as a replacement for time-oriented heuristics in practical web analytics of activity sessions.

In this work, we will challenge the assertion by prior works that (1) no reasonable cutoff is identifiable from the empirical data and (2) a global inactivity threshold is inappropriate as a session identification strategy. To our knowledge, we are the first to apply a general session identification methodology to a large collection of datasets and conclude that not only are global inactivity thresholds an appropriate strategy for session identification, but also that, for most user-initiated actions, an inactivity threshold of 1 hour is appropriate.

3. METHODS

This section is intended to both serve as a description of our methodology as well as to instruct readers on how to apply the same methods to their own datasets. First, we will discuss how we recommend applying our methodology for identifying inter-activity type component clusters to a dataset. Then we describe the origin of our datasets and the cleanup we performed in order to remove artifacts.

3.1 Fitting inter-activity times

First, we must gather a dataset of user-initiated actions with timestamps of at least *seconds* resolution. We generate inter-activity times on a per-user basis, so a relatively robust user identifier is necessary. While a persistent user identifier such as one associated with a user account is preferable, we have found that in the case of request logs, a fingerprint based on the request’s IP and User-agent seems to be sufficient.

Once we have generated per-user inter-activity times, we plot a histogram based on the logarithmically scaled inter-activity time and look for evidence of a valley. Given the observations we have seen (and report in section 4), we expect to see a valley around about 1 hour with peaks around 1 minute and 1 day. It is at this time that anomalies in the data should be detected and removed. For example, we found that the time between Wikimedia Mobile Views (de-

²Note that this conclusion was reached using the same AOL search dataset that we analyze in this paper

scribed in the next section) had an absurd spike at exactly 18 minutes of inter-activity time caused by a few (likely automated) users and removed their activities from the dataset.

Next, we try to fit a two component gaussian mixture model using expectation maximization [2] and visually inspect the results³. When the simple bimodal components did not appear to fit the data appropriately, we explored the addition of components to the mixture model with careful skepticism and repeated visual inspection.

Finally, if we have found what appears to be an appropriate fit, we identify a theoretically optimal inter-activity threshold for identifying sessions by finding the point where inter-activity time is equally likely to be within the gaussians fit with sub-hour means (within-session) and gaussians fit with means beyond an hour (between-session).

3.2 Datasets

To test this approach to session identification, we used a variety of datasets covering multiple sites, user groups and types of action.

Wikimedia sites. One of the broadest groups of datasets comes from the Wikimedia websites (such as Wikipedia) and covers both page views (read actions) and edits. For the page views, we gather three datasets, each consisting of randomly-sampled page view events from the Wikimedia request logs. These covered app views (page views from the Wikimedia’s official mobile app), mobile views (page views to the mobile site) and desktop views (page views to the desktop site). 100,000 IP addresses (or UIDs, in the case of the app, since it has those built in) were selected, and all requests from those IPs/UIDs for the month of October 2014 were retrieved. For desktop and mobile views, a UID was produced by hashing the IP address, the User agent, and the accept_language provided with each request. After filtering out known crawlers and automata using *tobie’s* *ua-parser*⁴, we arrived at three page view datasets consisting of 2,376,891, 932,754 and 2,285,521 pageviews, respectively. These came from 100,000, 235,067 or 247,269 UIDs. We also extracted inter-edit times from the English Wikipedia using the methodology we employed in [9] – randomly selecting 1 million edits from 157,342 registered users.

AOL search. Contrasting with the Wikimedia datasets we used the (now infamous) AOL search logs⁵ (*aol_search*) consisting of 36,389,567 search actions from 657,427 unique IDs. These actions span from March through May of 2006.

Cyclopath. We also gathered a dataset from Cyclopath, a computational geowiki leveraging cyclist knowledge [18]. The dataset consists of HTTP requests to the Cyclopath server that are automatically labelled by type. We filtered

³Note that we tried several strategies for statistically confirming the most appropriate fit – of which we found Davies-Bouldin index (DBI) [6] to be most reasonable – but none were as good as a simple visual inspection, so we employ and recommend the same.

⁴<https://github.com/tobie/ua-parser>

⁵These logs are controversial due to their inclusion of search terms containing private information, and there has historically been an ethical debate about their use. We are confident, however, that our usage does not have ethical implications; we modified the dataset to strip search terms so that it consists solely of unique IDs and timestamps, as has been used in the past [13]. See https://en.wikipedia.org/wiki/AOL_search_data_leak for more discussion.



https://commons.wikimedia.org/wiki/File:See_No_Evil_Hear_No_Evil_Speak_No_Evil.jpg CC-BY-SA 3.0

that 30 minutes produced artefacts around long sessions, and could find no clear evidence of a global session inactivity threshold², while Jones & Klinkner [12] found the 25.5 minute threshold performed “no better than random” in the context of identifying search tasks. Other thresholds have been proposed, but Montgomery and Faloutsos [14] concluded that the actual threshold chosen made little difference to how accurately sessions were identified.

Navigation-oriented heuristics involve inferring browsing patterns based on the HTTP referrers and URLs associated with each request by a user. When a user begins navigating (without a referrer), they have started a session; when a trail can no longer be traced to a previous request based on the referrers and URLs of subsequent requests, the session has ended. This approach was pioneered by Cooley et al in 2002 [5]. While it demonstrated utility in identifying “tasks”, and has been extended by Najjarbashi-Noghani et al. [15], it shows poor performance on sites with framesets due to implicit assumptions about web architecture [3]. Further, the sheer complexity of this strategy and it’s developmental focus on task over session make it unsuitable as a replacement for time-oriented heuristics in practical web analytics of activity sessions.

In this work, we will challenge the assertion by prior works that (1) no reasonable cutoff is identifiable from the empirical data and (2) a global inactivity threshold is inappropriate as a session identification strategy. To our knowledge, we are the first to apply a general session identification methodology to a large collection of datasets and conclude that not only are global inactivity thresholds an appropriate strategy for session identification, but also that, for most user-initiated actions, an inactivity threshold of 1 hour is appropriate.

3. METHODS

This section is intended to both serve as a description of our methodology as well as to instruct readers on how to apply the same methods to their own datasets. First, we will discuss how we recommend applying our methodology for identifying inter-activity type component clusters to a dataset. Then we describe the origin of our datasets and the cleanup we performed in order to remove artifacts.

3.1 Fitting inter-activity times

First, we must gather a dataset of user-initiated actions with timestamps of at least *seconds* resolution. We generate inter-activity times on a per-user basis, so a relatively robust user identifier is necessary. While a persistent user identifier such as one associated with a user account is preferable, we have found that in the case of request logs, a fingerprint based on the request’s IP and User-agent seems to be sufficient.

Once we have generated per-user inter-activity times, we plot a histogram based on the logarithmically scaled inter-activity time and look for evidence of a valley. Given the observations we have seen (and report in section 4), we expect to see a valley around about 1 hour with peaks around 1 minute and 1 day. It is at this time that anomalies in the data should be detected and removed. For example, we found that the time between Wikimedia Mobile Views (de-

²Note that this conclusion was reached using the same AOL search dataset that we analyze in this paper

scribed in the next section) had an absurd spike at exactly 18 minutes of inter-activity time caused by a few (likely automated) users and removed their activities from the dataset.

Next, we try to fit a two component gaussian mixture model using expectation maximization [2] and visually inspect the results³. When the simple bimodal components did not appear to fit the data appropriately, we explored the addition of components to the mixture model with careful skepticism and repeated visual inspection.

Finally, if we have found what appears to be an appropriate fit, we identify a theoretically optimal inter-activity threshold for identifying sessions by finding the point where inter-activity time is equally likely to be within the gaussians fit with sub-hour means (within-session) and gaussians fit with means beyond an hour (between-session).

3.2 Datasets

To test this approach to session identification, we used a variety of datasets covering multiple sites, user groups and types of action.

Wikimedia sites. One of the broadest groups of datasets comes from the Wikimedia websites (such as Wikipedia) and covers both page views (read actions) and edits. For the page views, we gather three datasets, each consisting of randomly-sampled page view events from the Wikimedia request logs. These covered app views (page views from the Wikimedia’s official mobile app), mobile views (page views to the mobile site) and desktop views (page views to the desktop site). 100,000 IP addresses (or UUIDs, in the case of the app, since it has those built in) were selected, and all requests from those IPs/UUIDs for the month of October 2014 were retrieved. For desktop and mobile views, a UUID was produced by hashing the IP address, the User agent, and the accept_language provided with each request. After filtering out known crawlers and automata using *tobie’s* *ua-parser*⁴, we arrived at three page view datasets consisting of 2,376,891, 932,754 and 2,285,521 pageviews, respectively. These came from 100,000, 235,067 or 247,269 UUIDs. We also extracted inter-edit times from the English Wikipedia using the methodology we employed in [9] – randomly selecting 1 million edits from 157,342 registered users.

AOL search. Contrasting with the Wikimedia datasets we used the (now infamous) AOL search logs⁵ (aol_search) consisting of 36,389,567 search actions from 657,427 unique IDs. These actions span from March through May of 2006.

Cyclopath. We also gathered a dataset from Cyclopath, a computational geowiki leveraging cyclist knowledge [18]. The dataset consists of HTTP requests to the Cyclopath server that are automatically labelled by type. We filtered

³Note that we tried several strategies for statistically confirming the most appropriate fit – of which we found Davies-Bouldin index (DBI) [6] to be most reasonable – but none were as good as a simple visual inspection, so we employ and recommend the same.

⁴<https://github.com/tobie/ua-parser>

⁵These logs are controversial due to their inclusion of search terms containing private information, and there has historically been an ethical debate about their use. We are confident, however, that our usage does not have ethical implications; we modified the dataset to strip search terms so that it consists solely of unique IDs and timestamps, as has been used in the past [13]. See https://en.wikipedia.org/wiki/AOL_search_data_leak for more discussion.



https://commons.wikimedia.org/wiki/File:Bernard_Trebacz_Argument_of_the_scholars.jpg Public Domain



https://commons.wikimedia.org/wiki/File:See_No_Evil,_Hear_No_Evil,_Speak_No_Evil.jpg CC-BY-SA 3.0

that 30 minutes produced artefacts around long sessions, and could find no clear evidence of a global session inactivity threshold², while Jones & Klinkner [12] found the 25.5 minute threshold performed “no better than random” in the context of identifying search tasks. Other thresholds have been proposed, but Montgomery and Faloutsos [14] concluded that the actual threshold chosen made little difference to how accurately sessions were identified.

Navigation-oriented heuristics involve inferring browsing patterns based on the HTTP referrers and URLs associated with each request by a user. When a user begins navigating (without a referrer), they have started a session; when a trail can no longer be traced to a previous request based on the referrers and URLs of subsequent requests, the session has ended. This approach was pioneered by Cooley et al in 2002 [5]. While it demonstrated utility in identifying “tasks”, and has been extended by Najjarbashi-Noghani et al. [15], it shows poor performance on sites with framesets due to implicit assumptions about web architecture [3]. Further, the sheer complexity of this strategy and it’s developmental focus on task over session make it unsuitable as a replacement for time-oriented heuristics in practical web analytics of activity sessions.

In this work, we will challenge the assertion by prior works that (1) no reasonable cutoff is identifiable from the empirical data and (2) a global inactivity threshold is inappropriate as a session identification strategy. To our knowledge, we are the first to apply a general session identification methodology to a large collection of datasets and conclude that not only are global inactivity thresholds an appropriate strategy for session identification, but also that, for most user-initiated actions, an inactivity threshold of 1 hour is appropriate.

3. METHODS

This section is intended to both serve as a description of our methodology as well as to instruct readers on how to apply the same methods to their own datasets. First, we will discuss how we recommend applying our methodology for identifying inter-activity type component clusters to a dataset. Then we describe the origin of our datasets and the cleanup we performed in order to remove artifacts.

3.1 Fitting inter-activity times

First, we must gather a dataset of user-initiated actions with timestamps of at least *seconds* resolution. We generate inter-activity times on a per-user basis, so a relatively robust user identifier is necessary. While a persistent user identifier such as one associated with a user account is preferable, we have found that in the case of request logs, a fingerprint based on the request’s IP and User-agent seems to be sufficient.

Once we have generated per-user inter-activity times, we plot a histogram based on the logarithmically scaled inter-activity time and look for evidence of a valley. Given the observations we have seen (and report in section 4), we expect to see a valley around about 1 hour with peaks around 1 minute and 1 day. It is at this time that anomalies in the data should be detected and removed. For example, we found that the time between Wikimedia Mobile Views (de-

²Note that this conclusion was reached using the same AOL search dataset that we analyze in this paper

scribed in the next section) had an absurd spike at exactly 18 minutes of inter-activity time caused by a few (likely automated) users and removed their activities from the dataset.

Next, we try to fit a two component gaussian mixture model using expectation maximization [2] and visually inspect the results³. When the simple bimodal components did not appear to fit the data appropriately, we explored the addition of components to the mixture model with careful skepticism and repeated visual inspection.

Finally, if we have found what appears to be an appropriate fit, we identify a theoretically optimal inter-activity threshold for identifying sessions by finding the point where inter-activity time is equally likely to be within the gaussians fit with sub-hour means (within-session) and gaussians fit with means beyond an hour (between-session).

3.2 Datasets

To test this approach to session identification, we used a variety of datasets covering multiple sites, user groups and types of action.

Wikimedia sites. One of the broadest groups of datasets comes from the Wikimedia websites (such as Wikipedia) and covers both page views (read actions) and edits. For the page views, we gather three datasets, each consisting of randomly-sampled page view events from the Wikimedia request logs. These covered app views (page views from the Wikimedia’s official mobile app), mobile views (page views to the mobile site) and desktop views (page views to the desktop site). 100,000 IP addresses (or UUIDs, in the case of the app, since it has those built in) were selected, and all requests from those IPs/UUIDs for the month of October 2014 were retrieved. For desktop and mobile views, a UUID was produced by hashing the IP address, the User agent, and the accept_language provided with each request. After filtering out known crawlers and automata using *tobie’s* *ua-parser*⁴, we arrived at three page view datasets consisting of 2,376,891, 932,754 and 2,285,521 pageviews, respectively. These came from 100,000, 235,067 or 247,269 UUIDs. We also extracted inter-edit times from the English Wikipedia using the methodology we employed in [9] – randomly selecting 1 million edits from 157,342 registered users.

AOL search. Contrasting with the Wikimedia datasets we used the (now infamous) AOL search logs⁵ (aol, search) consisting of 36,389,567 search actions from 657,427 unique IDs. These actions span from March through May of 2006.

Cyclopath. We also gathered a dataset from Cyclopath, a computational geowiki leveraging cyclist knowledge [18]. The dataset consists of HTTP requests to the Cyclopath server that are automatically labelled by type. We filtered

³Note that we tried several strategies for statistically confirming the most appropriate fit – of which we found Davies-Bouldin index (DBI) [6] to be most reasonable – but none were as good as a simple visual inspection, so we employ and recommend the same.

⁴<https://github.com/tobie/ua-parser>

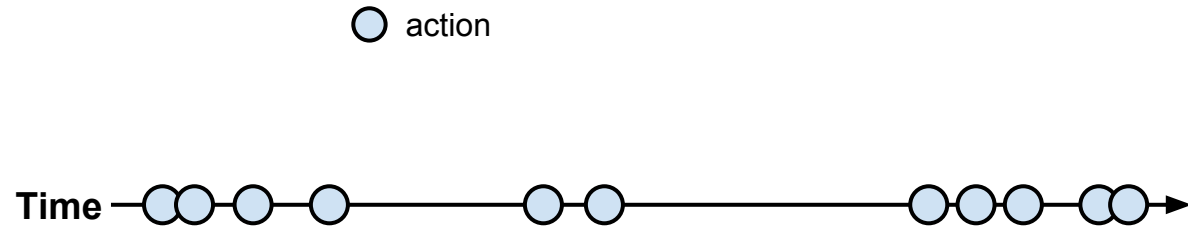
⁵These logs are controversial due to their inclusion of search terms containing private information, and there has historically been an ethical debate about their use. We are confident, however, that our usage does not have ethical implications; we modified the dataset to strip search terms so that it consists solely of unique IDs and timestamps, as has been used in the past [13]. See https://en.wikipedia.org/wiki/AOL_search_data_leak for more discussion.



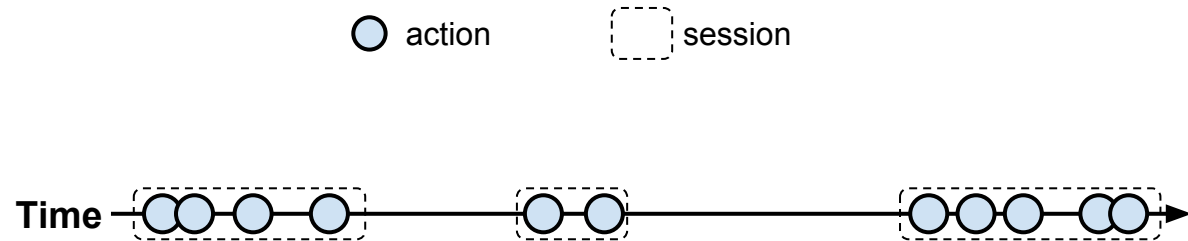
https://commons.wikimedia.org/wiki/File:Bernard_Trebacz_Argument_of_the_scholars.jpg Public Domain

“Methodology Nerd”

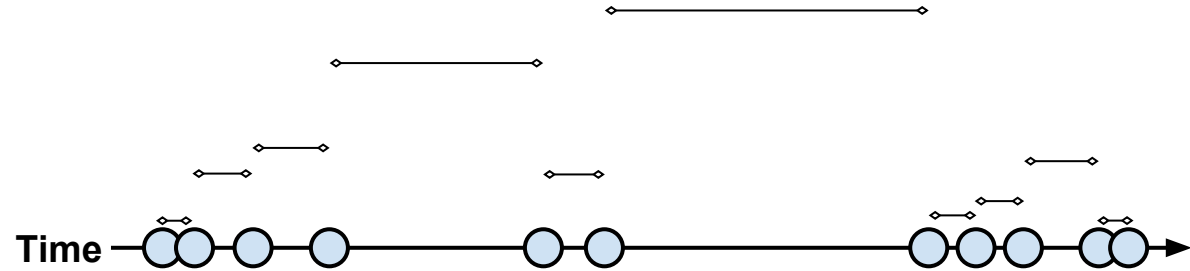
The problem: Session delineation



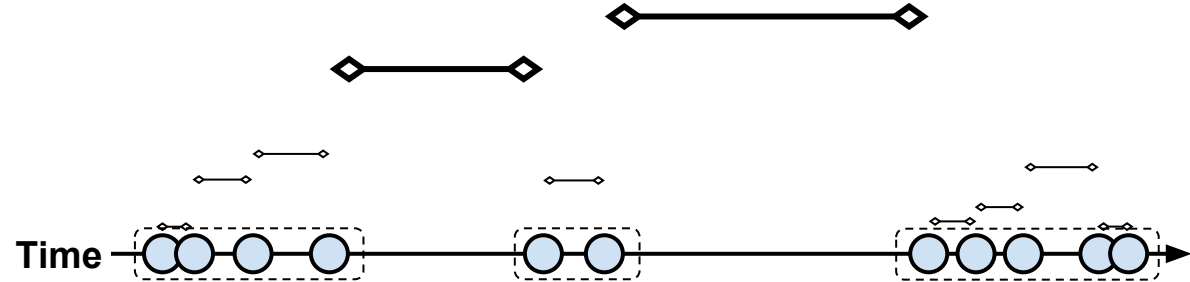
The problem: Session delineation



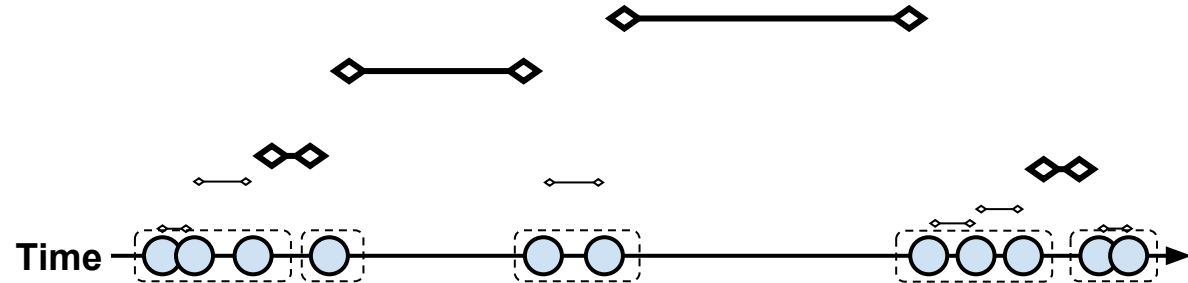
The problem: Session delineation



The problem: Session delineation



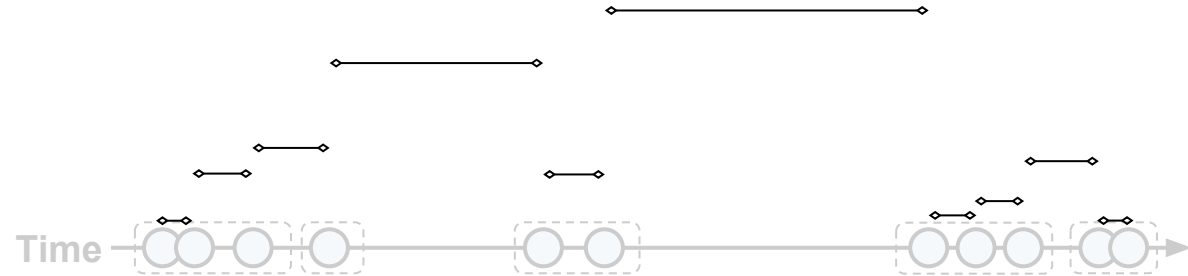
The problem: Session delineation



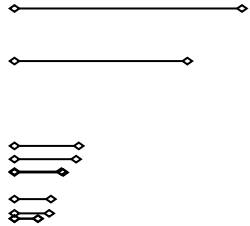
The problem: Session delineation

If “sessions” are a *meaningful* component of human activity, we should see a natural cutoff emerge in the data.

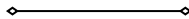
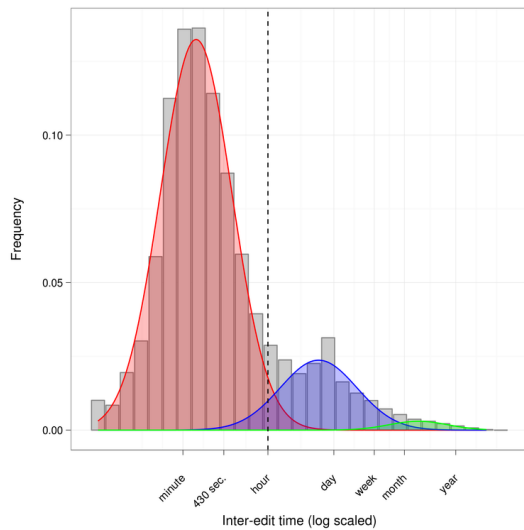
The problem: Session delineation

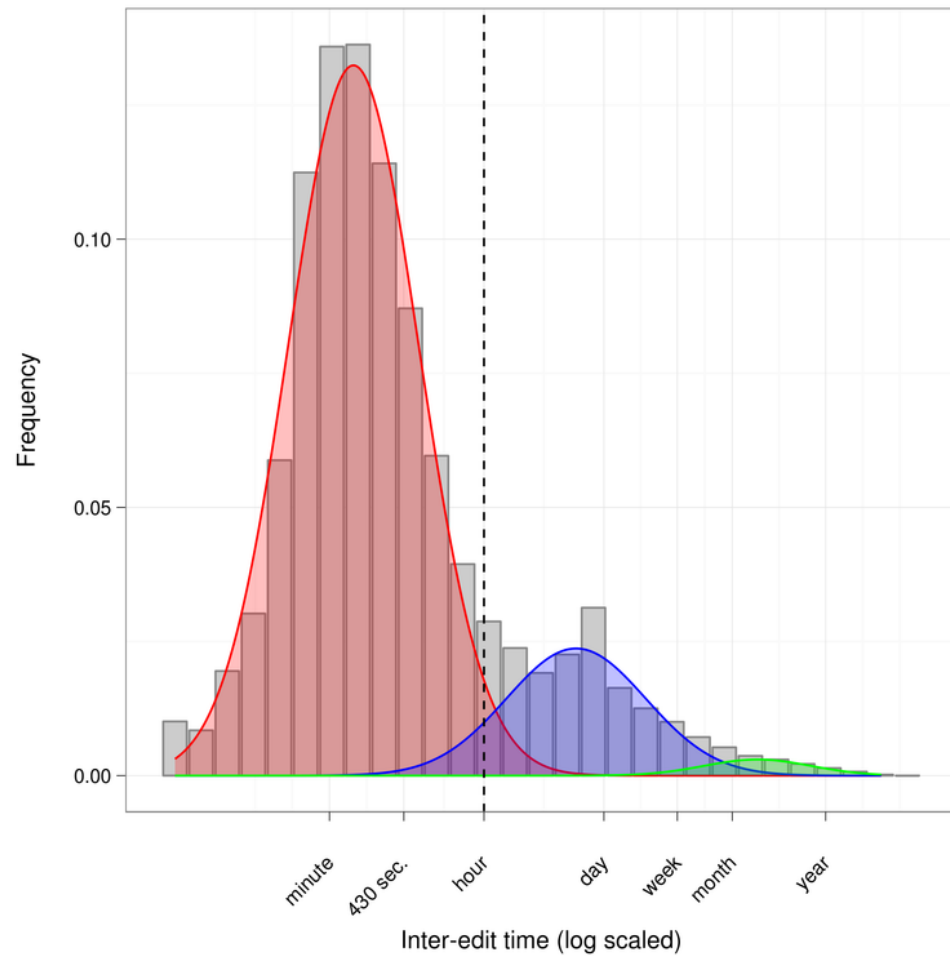
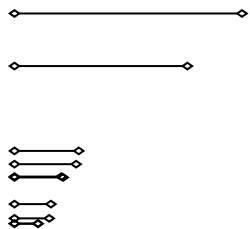


The problem: Session delineation

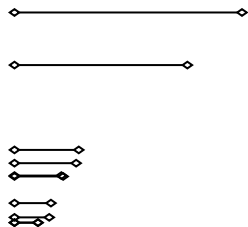
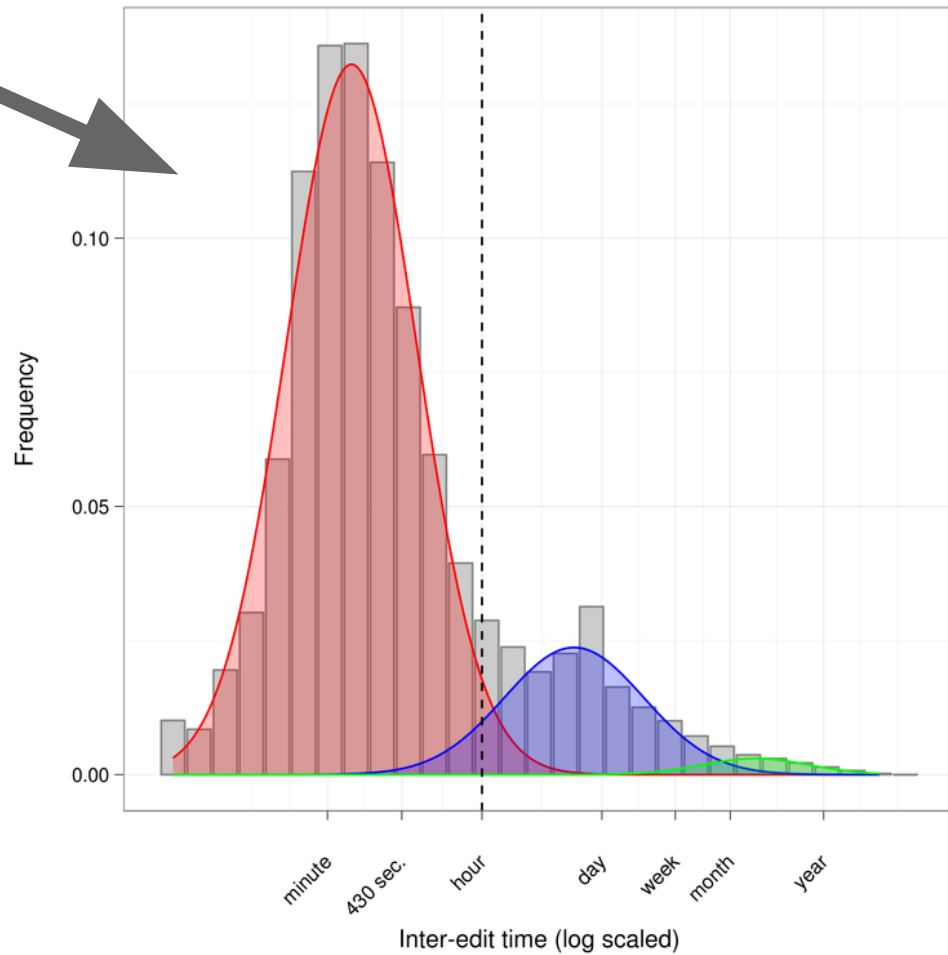


The problem: Session delineation





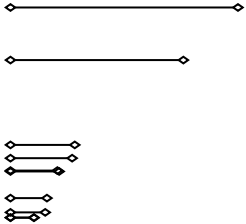
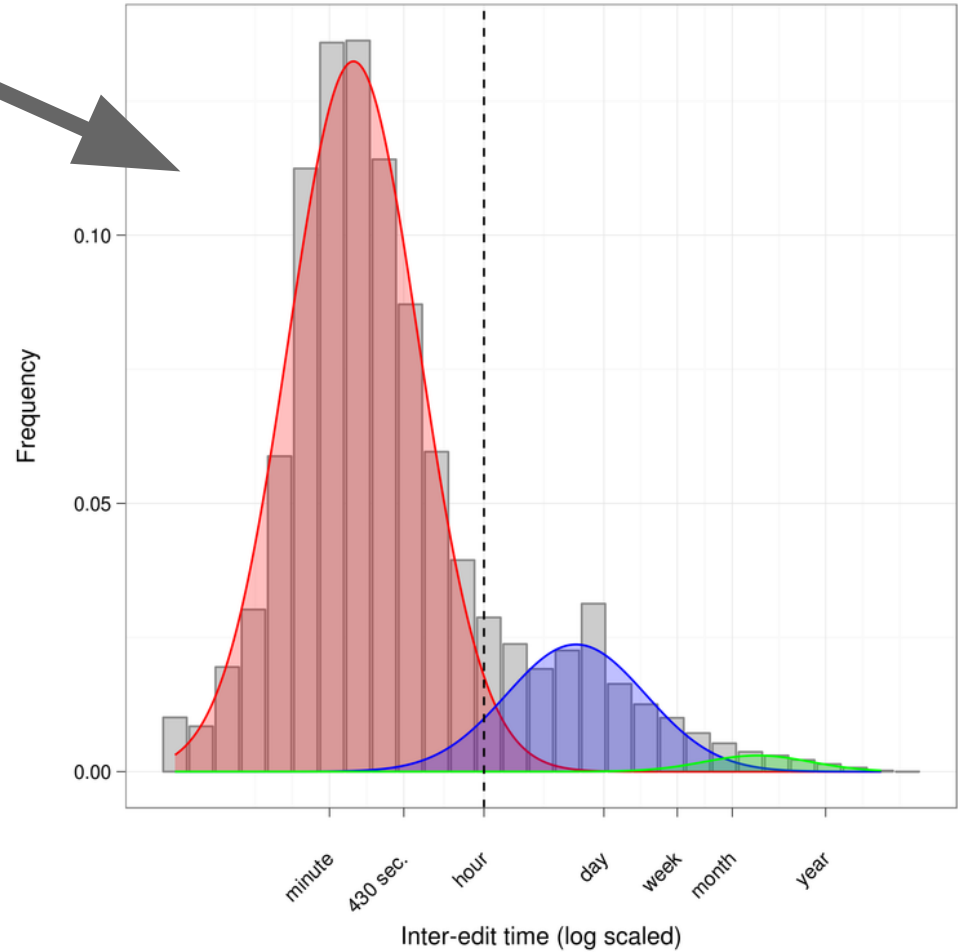
Expectation maximization fit
of a mixture model of
3 gaussians on log-scaled data

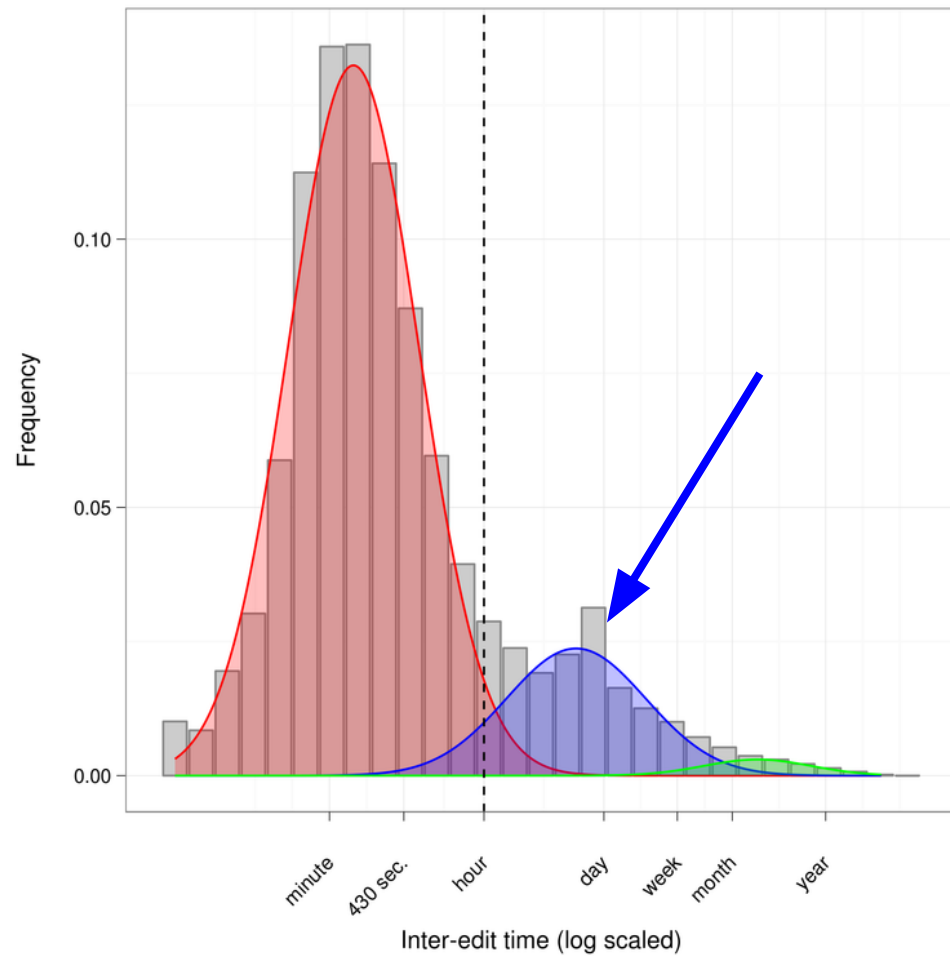
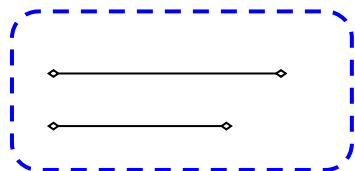


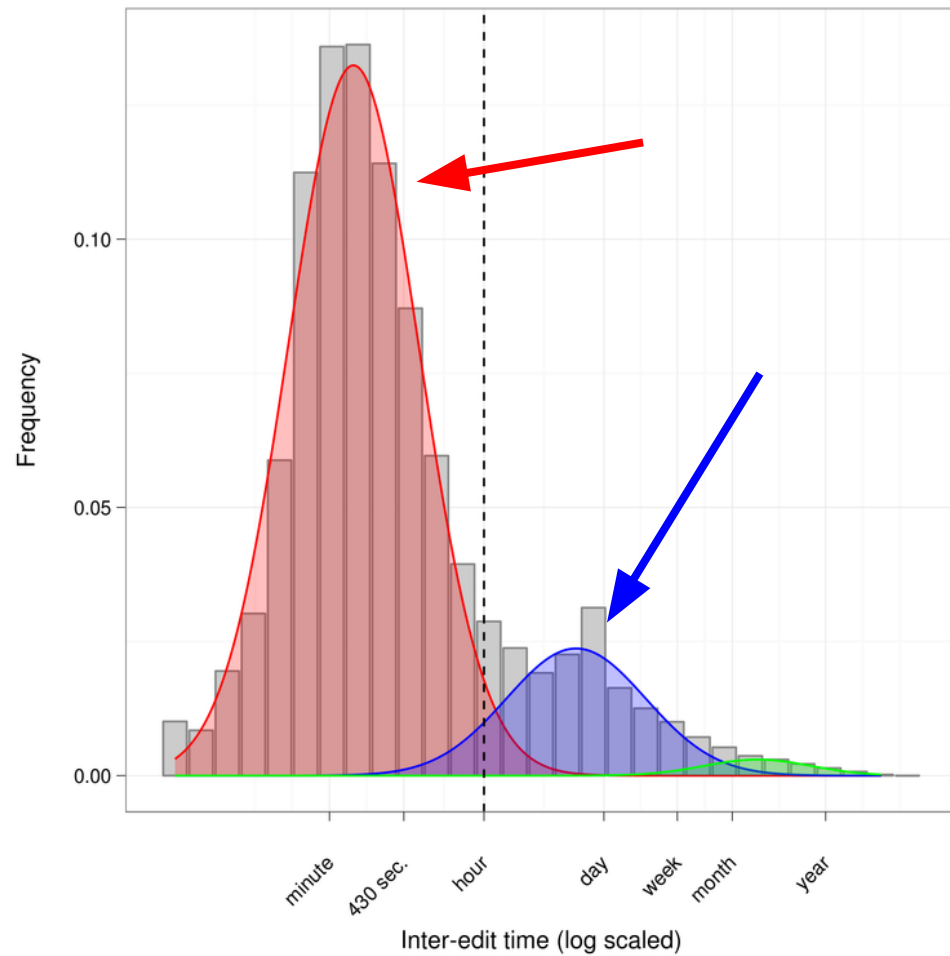
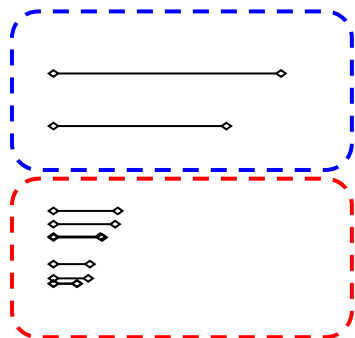
Expectation maximization fit
of a mixture model of
3 gaussians on log-scaled data

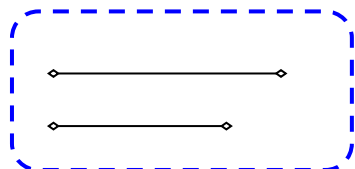


<http://arxiv.org/pdf/1411.2878v1.pdf>

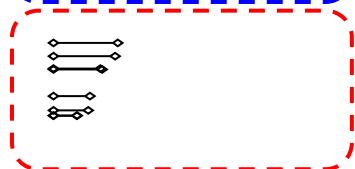




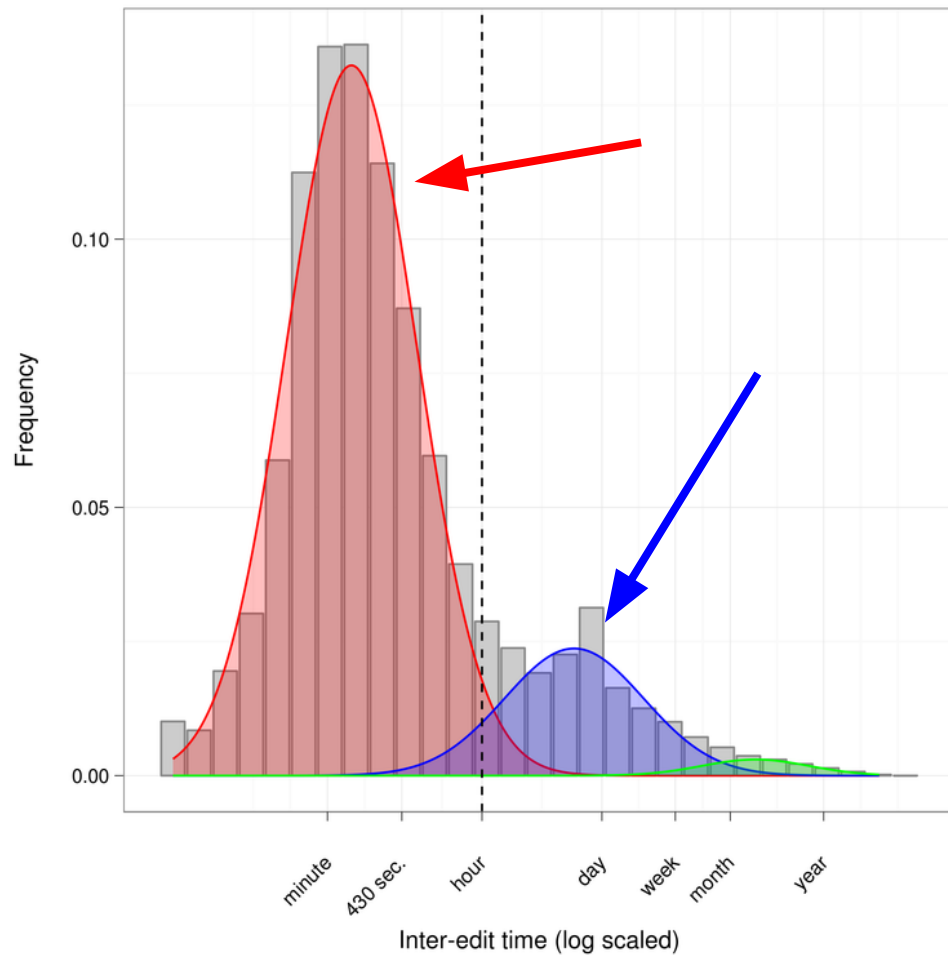


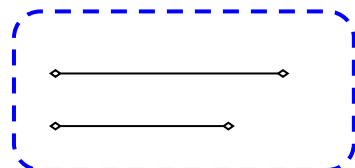


Between session
(12h - 1 week)

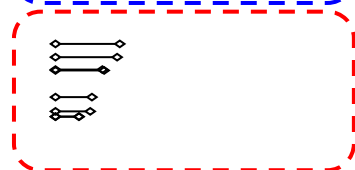


Within session
(1-7 minutes)

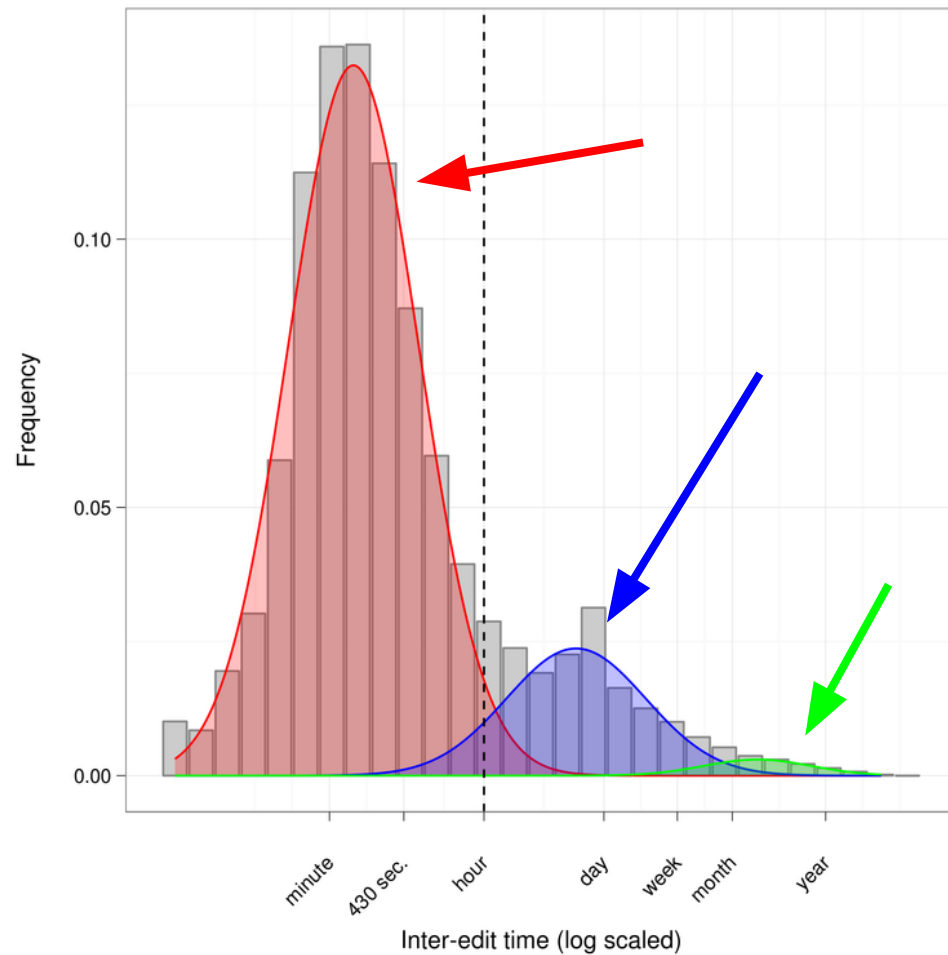


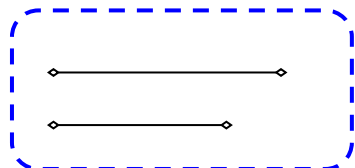


Between session
(12h - 1 week)



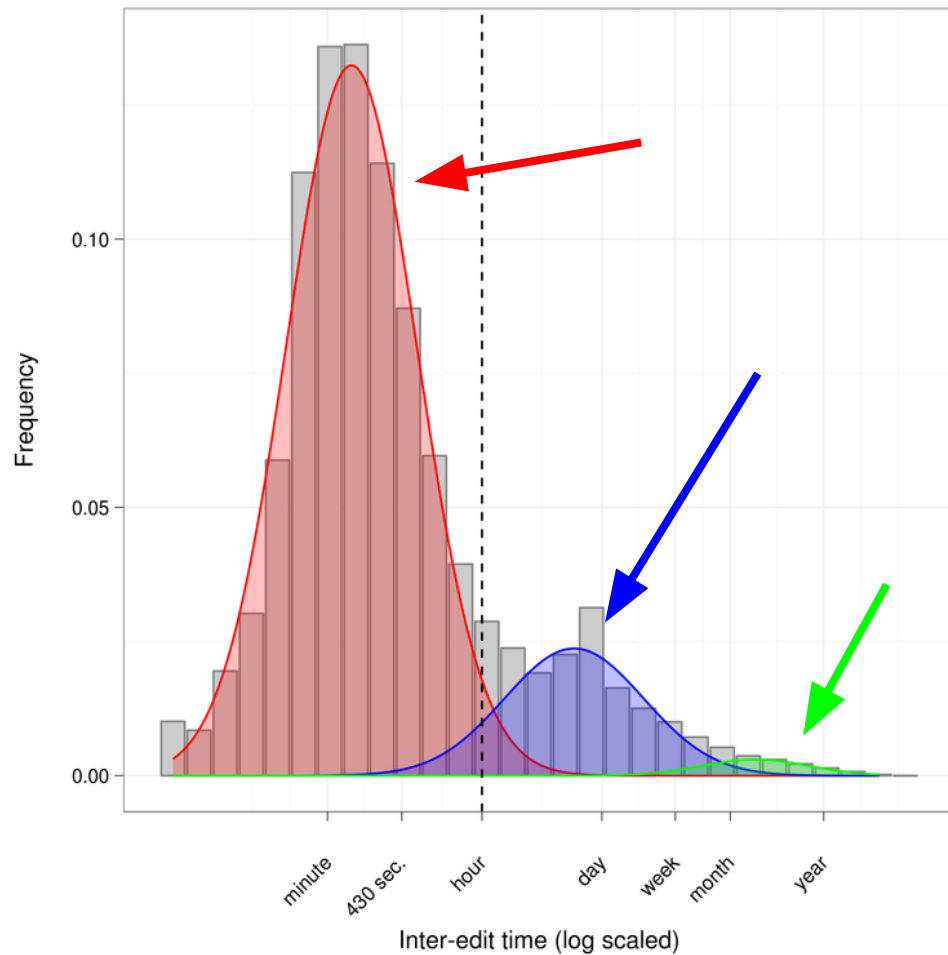
Within session
(1-7 minutes)





Between session
(12h - 1 week)

Within session
(1-7 minutes)



When to depart [\[edit\]](#)

- You are going to go **traveling**.
- You are a student and you have **exams** or other pressing schoolwork.
- You **marry** or **divorce**.
- You feel exhausted by your regional **weather**.
- You are neglecting your **work and job**.
- You get addicted to a **new video game** that you have to obsess over for the next few days.
- You find yourself repeatedly **assuming bad faith**.
- You are **stressed** to the point that it affects your editing.
- You have a new **girlfriend** or **boyfriend**.
- Your **girlfriend** or **boyfriend** wants you to spend more time with them.
- You **break up** with your boyfriend or girlfriend.
- You have to look for a **job**.
- You have a **job** and do not want to **lose it**.

When to depart [\[edit\]](#)

- You are going to go **traveling**.
- You are a student and you have **exams** or other pressing schoolwork.
- You **marry** or **divorce**.
- You feel exhausted by your regional **weather**.
- You are neglecting your **work and job**.

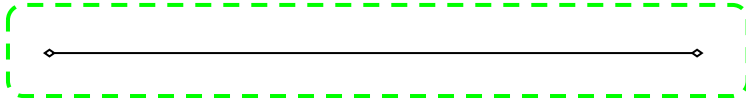
You get addicted to a n [When to come back](#) [\[edit\]](#)

- You find yourself repeat
- You are **stressed** to the
- You have a new **girlfriend**
- Your **girlfriend** or **boyfriend**
- You **break up** with your
- You have to look for a job
- You have a **job** and do
- You came back from travelling.
- Your relatives (thankfully) left your house.
- You have either been unblocked or your **block** has expired.
- Your spouse asks you to buy a new **car**. Just buy it and **come straight back**.
- Your **exams** are over.
- You decide that your **divorce** has left you more time to be a Wikipedian.
- You find out your new **girlfriend** or **boyfriend** is a Wikipedian, too.
- Especially when you find that now you've been spending a lot of time editing, discussing, fighting over and making up with said **friend**.
- You have finished **moving**.
- You decide to give up smoking so that you can pay your **Internet bill**.
- You've recovered from your **sickness**.
- You have been **reincarnated**.
- You realize you really **hadn't died** in the first place.

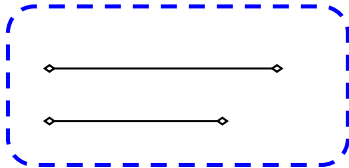
- You realize you really **hadn't died** in the first place.

-- WP:Wikibreak

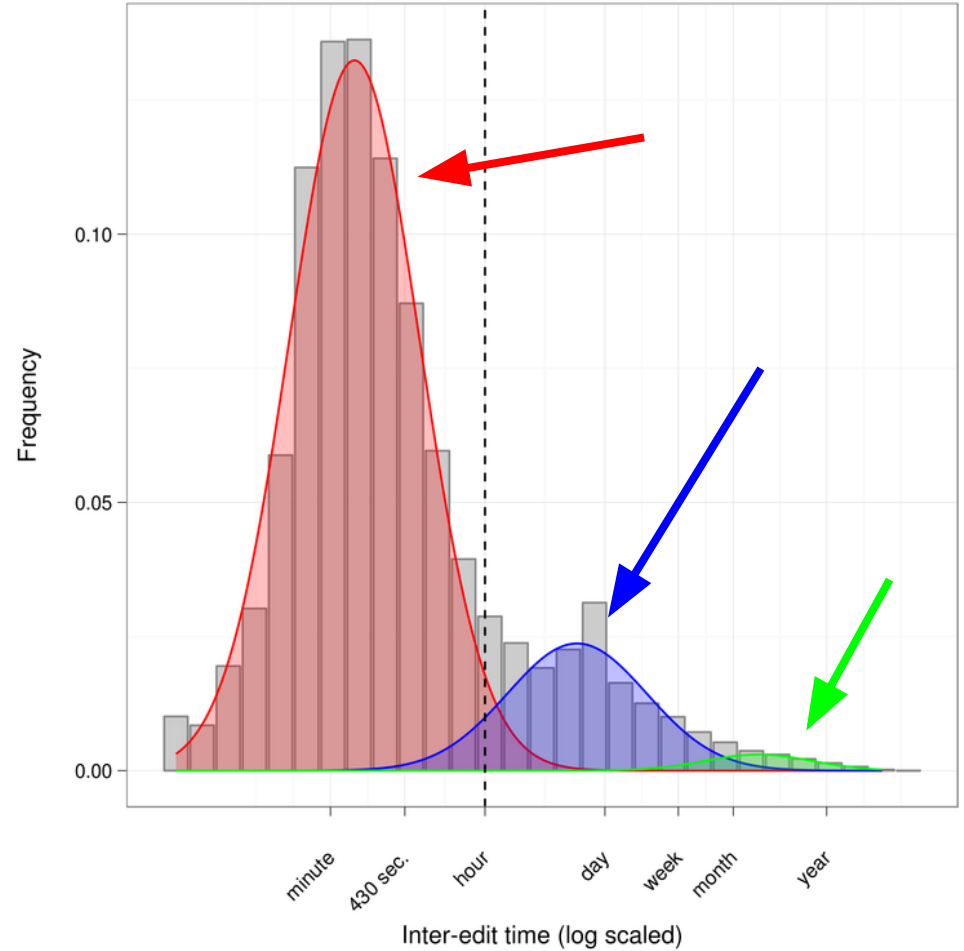
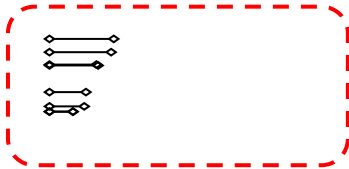
Break
(1-12 months)

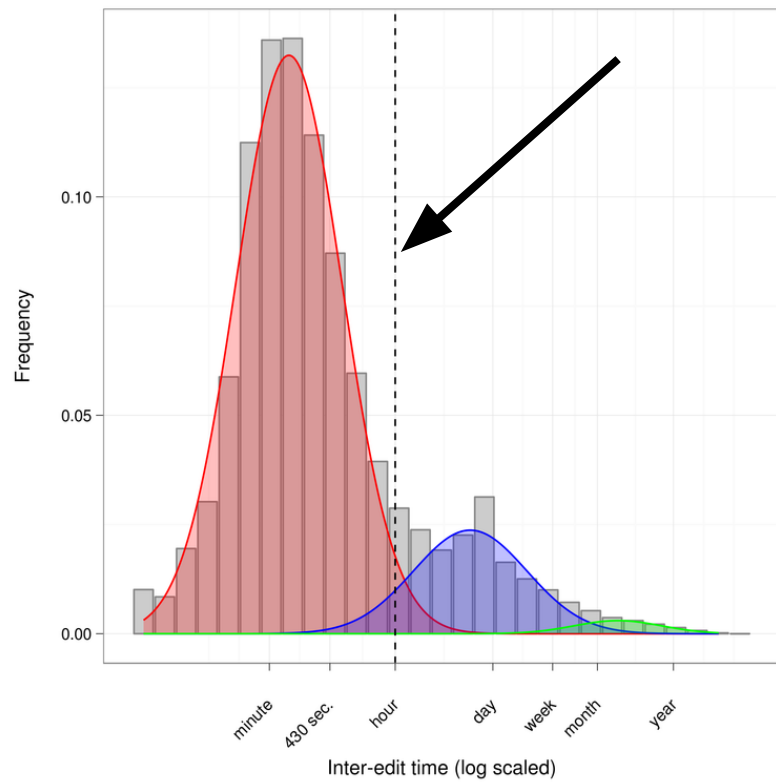


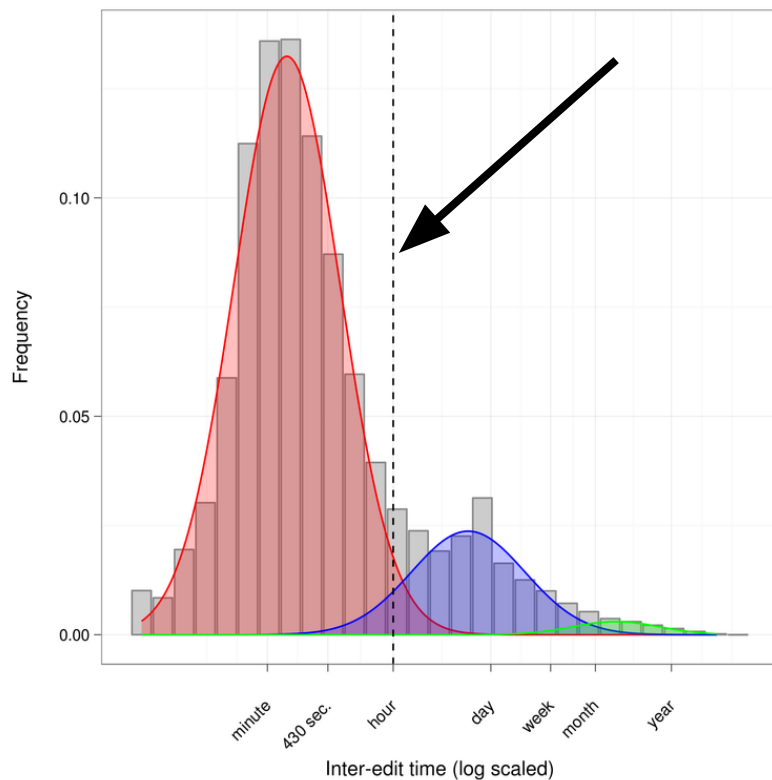
Between session
(12h - 1 week)



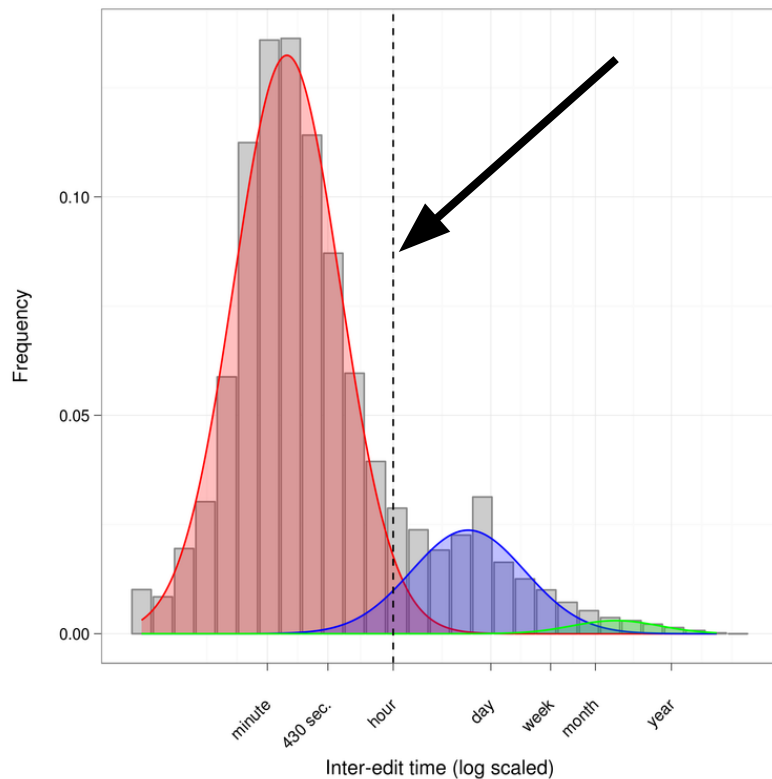
Within session
(1-7 minutes)



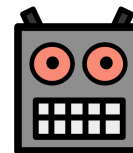


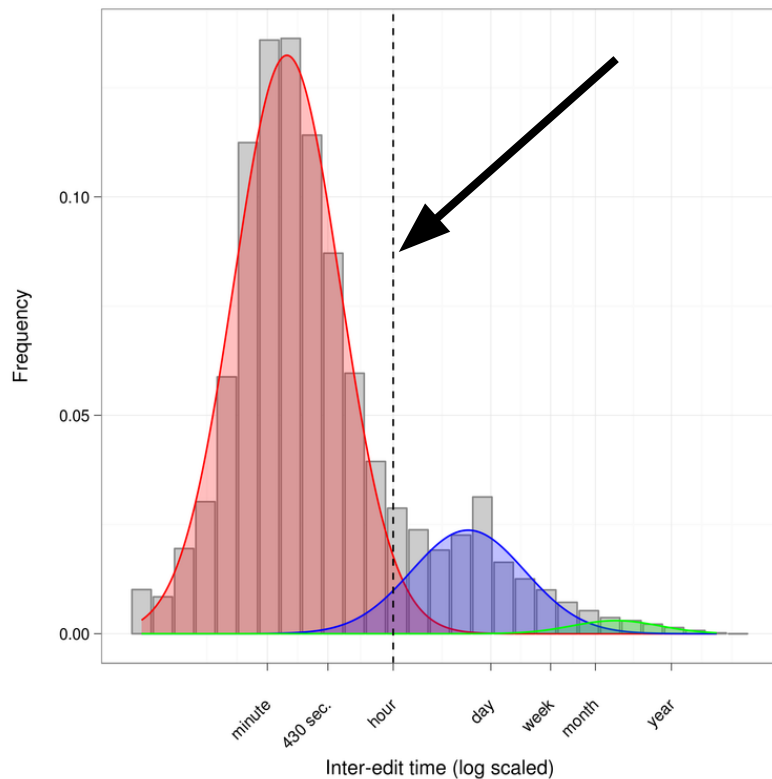


User	Revisions	User	Revisions	Hours
Koavf	43997	Materialscientist	7472	453.2664
Waacstats	33402	Kwamikagami	9088	356.3586
Hmains	17176	TonyTheTiger	6152	344.0164
Rich Farmbrough	17169	ACP2011	2218	337.2836
Bgwhite	14531	Pinethicket	3894	317.7897
Courcelles	13832	Armbrust	6288	311.0719
Fortdj33	12919	P.T. Aufrette	6257	306.4386
VasuVR	12095	Koavf	43997	302.2483
BD2412	9801	Derek R Bullamore	4228	290.0472
Cloudz679	9779	MathewTownsend	1807	280.8925
Kwamikagami	9088	Crisco 1492	2747	278.5694
Muboshgu	8098	Alarbus	1669	277.5844
Tassedethe	7976	Rich Farmbrough	17169	274.8825
Materialscientist	7472	Alan Liefting	5970	274.3742
John of Reading	7415	BD2412	9801	273.2358
DBigXray	7405	Sitush	4421	270.7494
Ssriram mt	7100	DBigXray	7405	270.2650
Woohookitty	7099	Allens	6757	270.1681
Allens	6757	Cloudz679	9779	249.9444

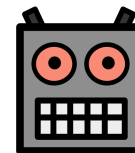


User	Revisions	User	Revisions	Hours
Koavf	43997	Materialscientist	7472	453.2664
Waacstats	33402	Kwamikagami	9088	356.3586
Hmains	17176	TonyTheTiger	6152	344.0164
Rich Farmbrough	17169	ACP2011	2218	337.2836
Bgwhite	14531	Pinethicket	3894	317.7897
Courcelles	13832	Armbrust	6288	311.0719
Fortdj33	12919	P.T. Aufrette	6257	306.4386
VasuVR	12095	Koavf	43997	302.2483
BD2412	9801	Derek R Bullamore	4228	290.0472
Cloudz679	9779	MathewTownsend	1807	280.8925
Kwamikagami	9088	Crisco 1492	2747	278.5694
Muboshgu	8098	Alarbus	1669	277.5844
Tassedethe	7976	Rich Farmbrough	17169	274.8825
Materialscientist	7472	Alan Liefting	5970	274.3742
John of Reading	7415	BD2412	9801	273.2358
DBigXray	7405	Sitush	4421	270.7494
Ssriram mt	7100	DBigXray	7405	270.2650
Woohookitty	7099	Allens	6757	270.1681
Allens	6757	Cloudz679	9779	249.9444





User	Revisions	User	Revisions	Hours
Koavf	43997	Materialscientist	7472	453.2664
Waacstats	33402	Kwamikagami	9088	356.3586
Hmains	17176	TonyTheTiger	6152	344.0164
Rich Farmbrough	17169	ACP2011	2218	337.2836
Bgwhite	14531	Pinethicket	3894	317.7897
Courcelles	13832	Armbrust	6288	311.0719
Fortdj33	12919	P.T. Aufrette	6257	306.4386
VasuVR	12095	Koavf	43997	302.2483
BD2412	9801	Derek R Bullamore	4228	290.0472
Cloudz679	9779	MathewTownsend	1807	280.8925
Kwamikagami	9088	Crisco 1492	2747	278.5694
Muboshgu	8098	Alarbus	1669	277.5844
Tassedethe	7976	Rich Farmbrough	17169	274.8825
Materialscientist	7472	Alan Liefting	5970	274.3742
John of Reading	7415	BD2412	9801	273.2358
DBigXray	7405	Sitush	4421	270.7494
Ssriram mt	7100	DBigXray	7405	270.2650
Woohookitty	7099	Allens	6757	270.1681
Allens	6757	Cloudz679	9779	249.9444



Channel tunnel

170 million

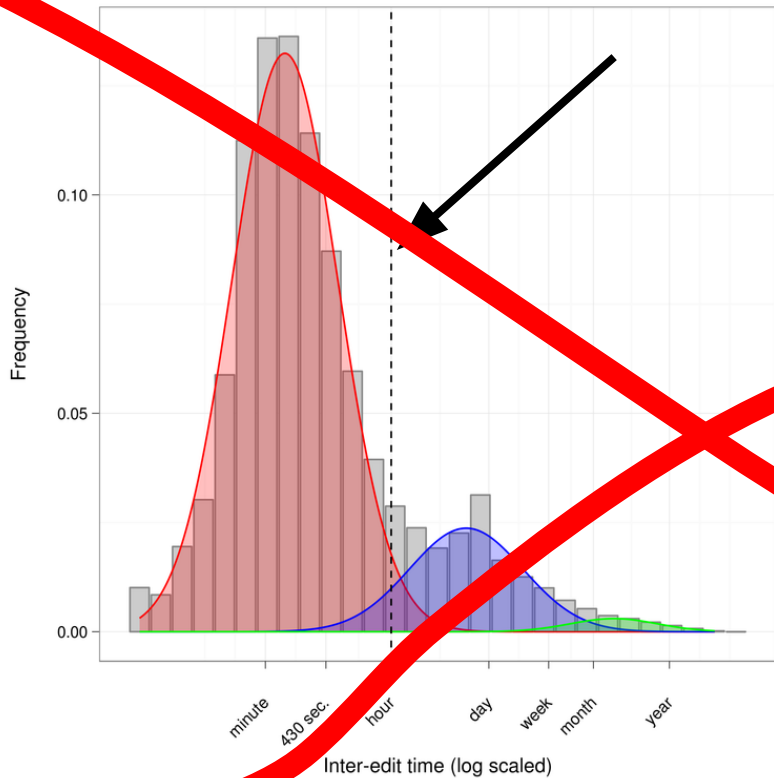
Wikipedia

140 million

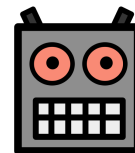
* based on 12M/year trend

Encyclopedia Britannica v1

12 thousand



User	Revisions	User	Revisions	Hours
Koavf	43997	Materialscientist	7472	453.2664
Waacstats	33402	Kwamikagami	9088	356.3586
Hmains	17176	TonyTheTiger	6152	344.0164
Rich Farmbrough	17169	ACP2011	2218	337.2836
Bgwhite	14531	Pineticket	3894	317.7897
Courcelles	13832	Arndt	6288	311.0719
Fortdj33	12919	M.T. Aufrette	6257	306.4386
VasuVR	12095	Koavf	43997	302.2483
BD2412	9779	Derek R Bullamore	4228	290.0472
Cloudz679	9779	MathewTownsend	1807	280.8925
Kwamikagami	9088	Crisco 1492	2747	278.5694
Muboshgu	8098	Alarbus	1669	277.5844
Tassedethe	7976	Rich Farmbrough	17169	274.8825
Materialscientist	7472	Alan Liefting	5970	274.3742
Pineticket Reading	7415	BD2412	9801	273.2358
DBigXray	7405	Sitush	4421	270.7494
Ssriram mt	7100	DBigXray	7405	270.2650
Woohookitty	7099	Allens	6757	270.1681
Allens	6757	Cloudz679	9779	249.9444



Channel tunnel

170 million

Wikipedia

140 million

* based on 12Myc

Encyclopedia Britannica v1

12 thousand

If this worked for **Wikipedia editing**, shouldn't it work elsewhere?

If this worked for **Wikipedia editing**, shouldn't it work elsewhere?

Oliver Keyes

Wikipedian

Let us fight to free the world, to do away with national barriers, to do away with greed, with hate and intolerance. Let us fight for a world of reason, a world where science and progress will lead to all men's happiness. ~ Charlie Chaplin



About me

I'm a long-term Wikipedian. In the real world I [work for the Wikimedia Foundation](#) as a researcher. Contributions under

My work

Mostly I focus on writing new articles and reviewing those created by other editors. Recently I've switched to trying to help out

Contact me

You can leave me a message on my [talk page](#). You can also email me at



- search



- desktop pageview
- mobile pageview
- app pageview
- edits



- route search



- search
- rate movie



- changeset



- ask question
- answer question



- game



- search



- desktop pageview
- mobile pageview
- app pageview
- edits



- route search

movielens
helping you find the *right* movies

- search
- rate movie



- changeset



- ask question
- answer question

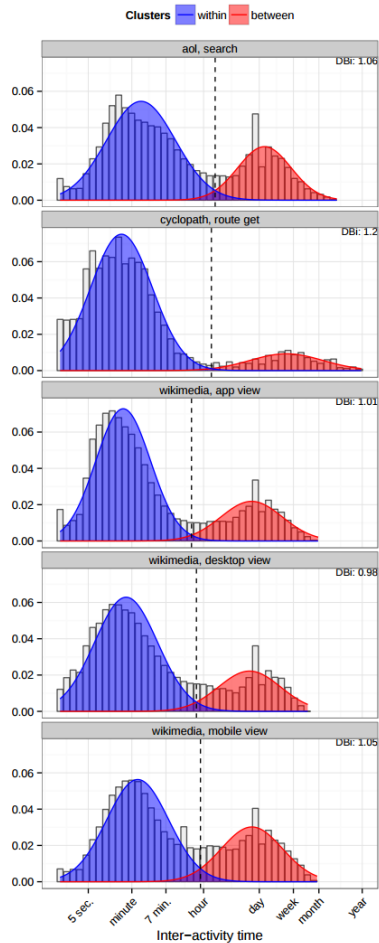


- game

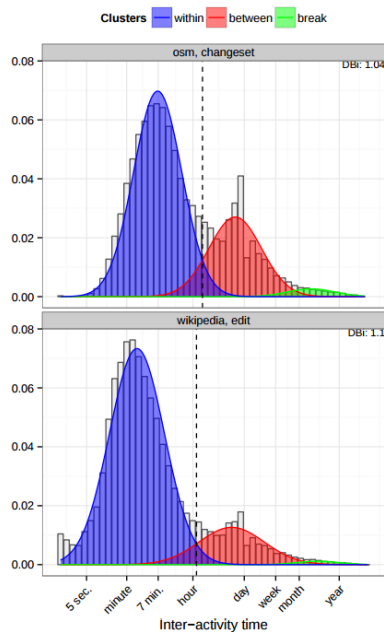
“MEGA-STUDY”

-- Dario Taraborelli

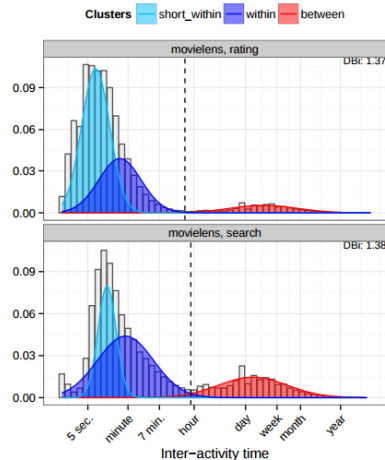
Bimodals



With break

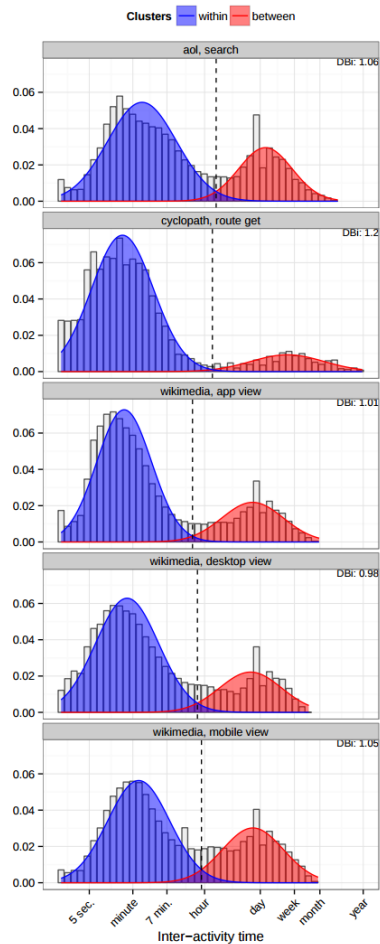


Short component

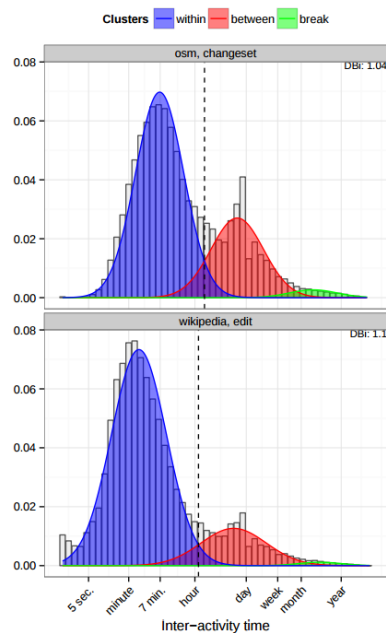


Strong regularity!

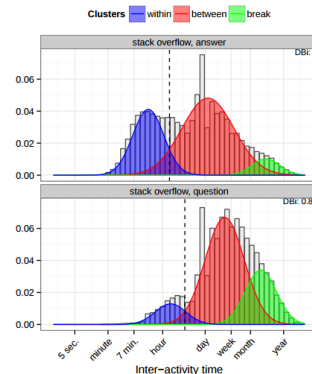
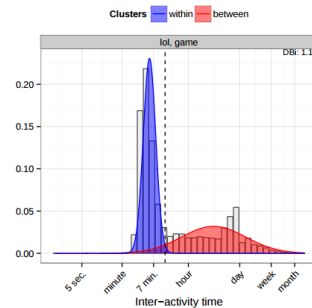
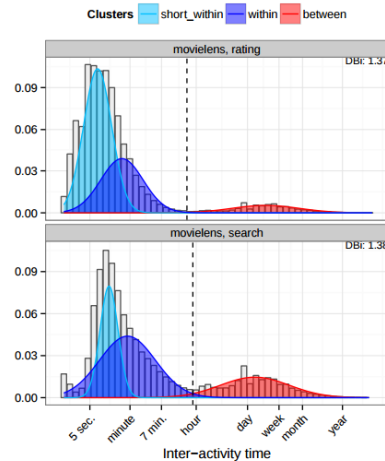
Bimodals



With break

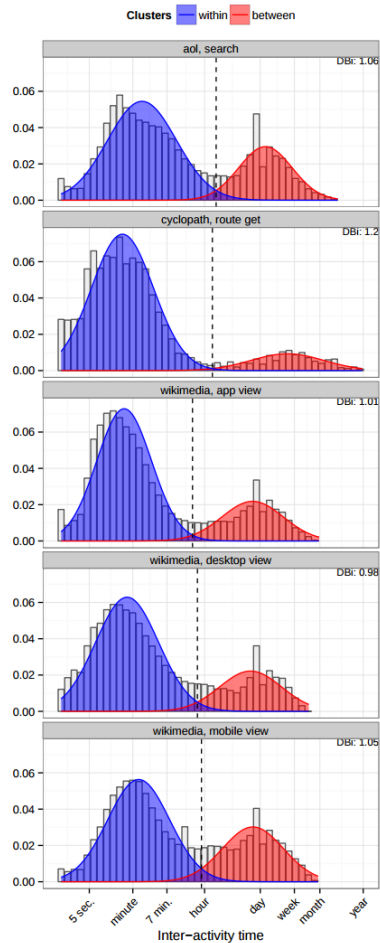


Short component

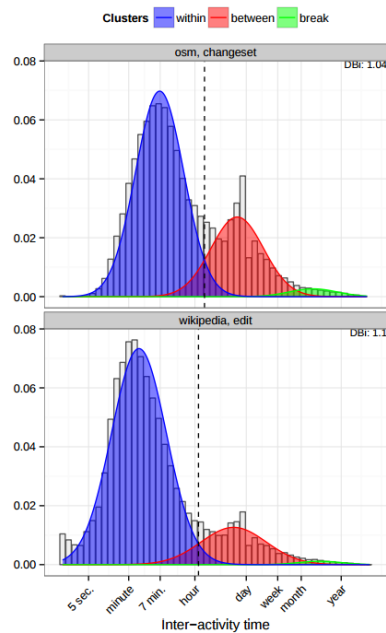


<http://arxiv.org/pdf/1411.2878v1.pdf>

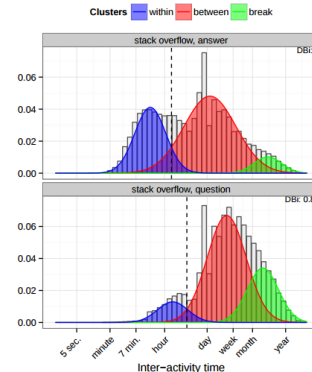
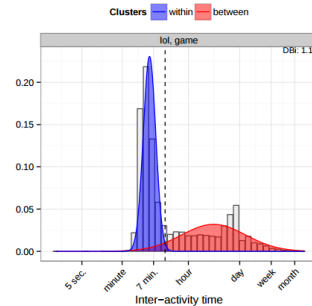
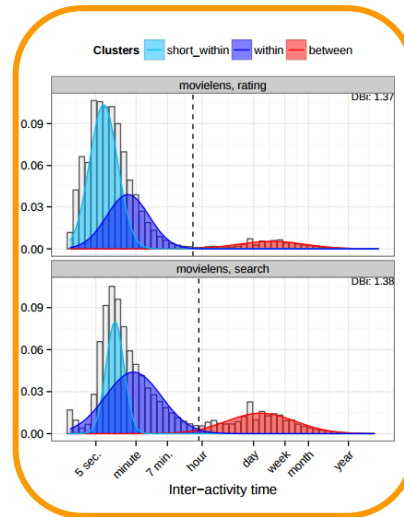
Bimodals



With break

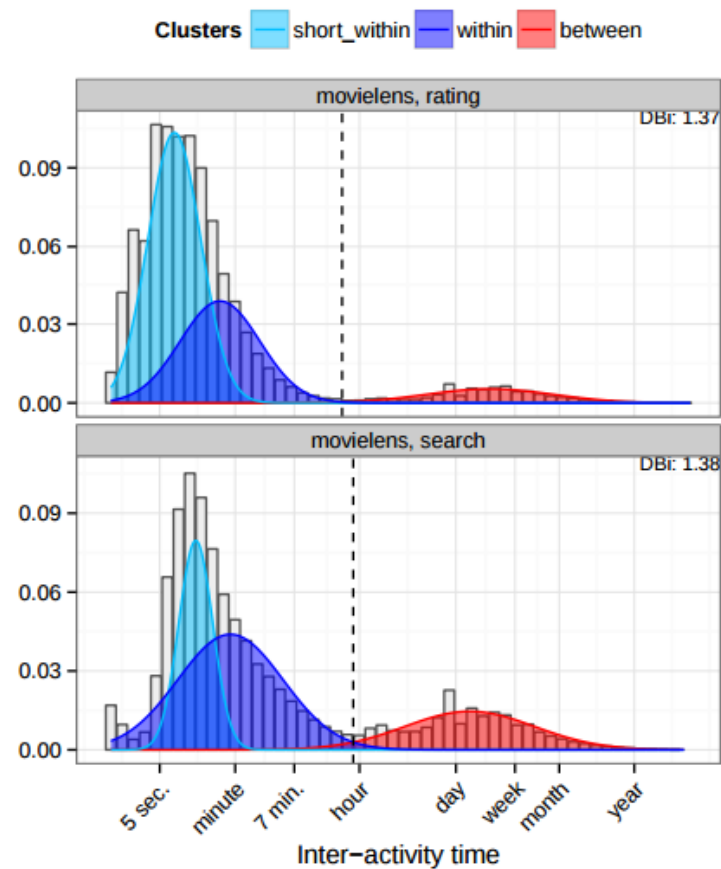


Short component



<http://arxiv.org/pdf/1411.2878v1.pdf>

Short component



Activity theory

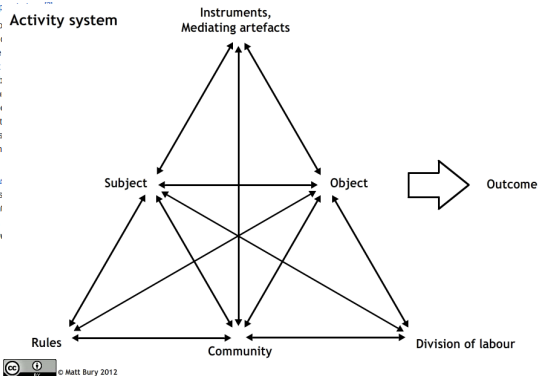
From Wikipedia, the free encyclopedia

For the psychosocial theory of aging, see *Activity theory (aging)*.

Activity theory (**AT**; Russian: *Теория деятельности*)^[1] is an umbrella term for a line of eclectic social sciences theories and research with its roots in the Soviet psychological activity theory pioneered by Lev Vygotsky, Alexei Leont'ev and Sergei Rubinstein. These scholars sought to understand human activities as complex, socially situated phenomena and to go beyond paradigms of reflexivity (the teaching of Vladimir Bekhterev and his followers) and *physiology of higher nervous activity* (the teaching of Ivan Pavlov and his school), *psychoanalysis* and *behaviorism*. It became one of the major psychological approaches in the former USSR, being widely used in both theoretical and applied psychology, and in education, professional training, ergonomics, social psychology and work psychology.

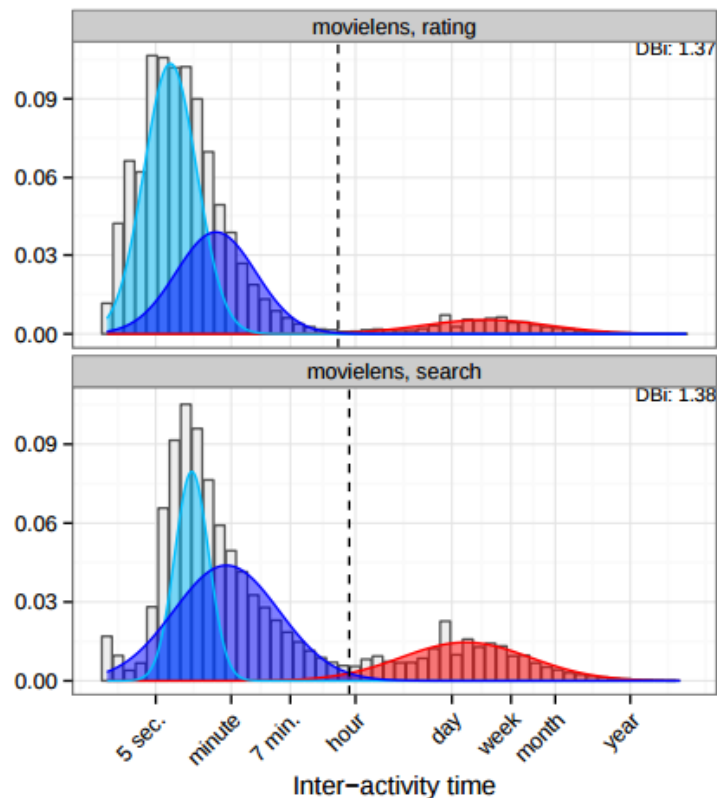
Activity theory is more of a descriptive meta-theory or framework than a predictive theory. It concerns organizations, etc.) beyond just one actor or user. It accounts for environment, history, motivations, and complexity of real life activity. One of the strengths of AT is that it bridges the reality—it studies both through the mediating activity. The unit of analysis in AT is the concept mediated human activity, or activity system. This system includes the object (or objective), subcommunity and division of labor. The motive for the activity in AT is created through the tension system.^[2] According to ethnographer Bonnie Nardi, a leading theorist in AT, activity theory "to distinguish 'applied' from 'pure' science—understanding everyday practice in the real world is the object of activity theory is to understand the unity of consciousness and activity."^[3] Sometimes approach is particularly useful for studying a group that exists "largely in virtual form, its common and printed texts."^[4]

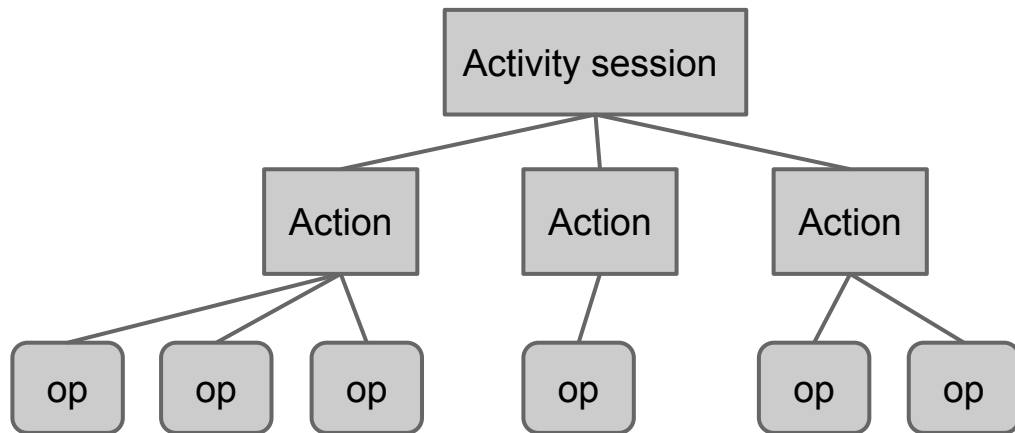
AT is particularly useful as a lens in qualitative research methodologies (e.g., *ethnography*, *case study*) and analyzing a phenomenon, finding patterns and making inferences across interactions, *descriptions* through a built-in language and rhetoric. A particular activity is a goal-directed or purposeful use of tools. These tools are externalized forms of mental processes manifested in constructs, the internalization and externalization of cognitive processes involved in the use of tools, as well as results from the interaction.^[5]



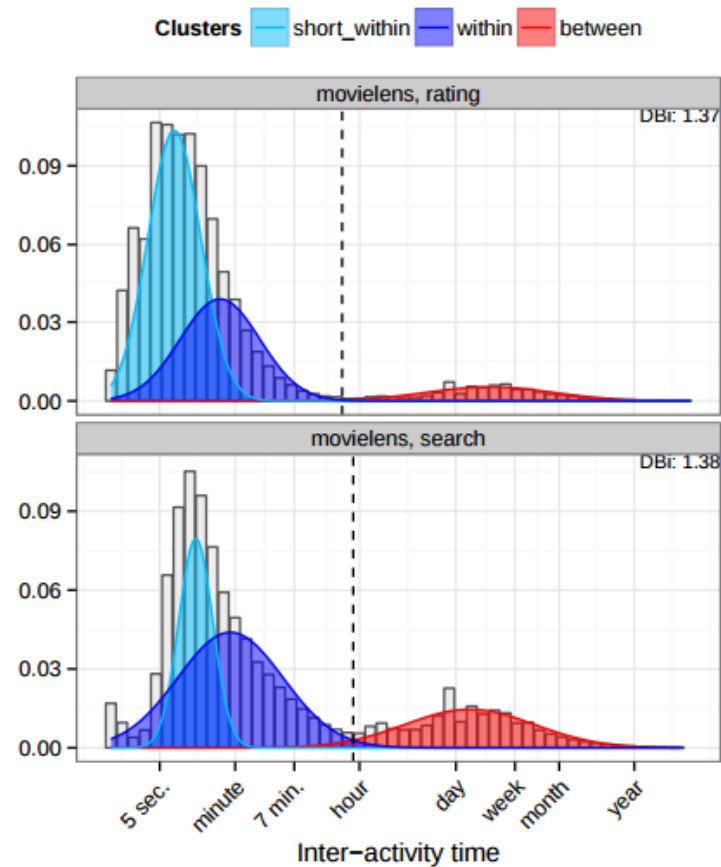
Short component

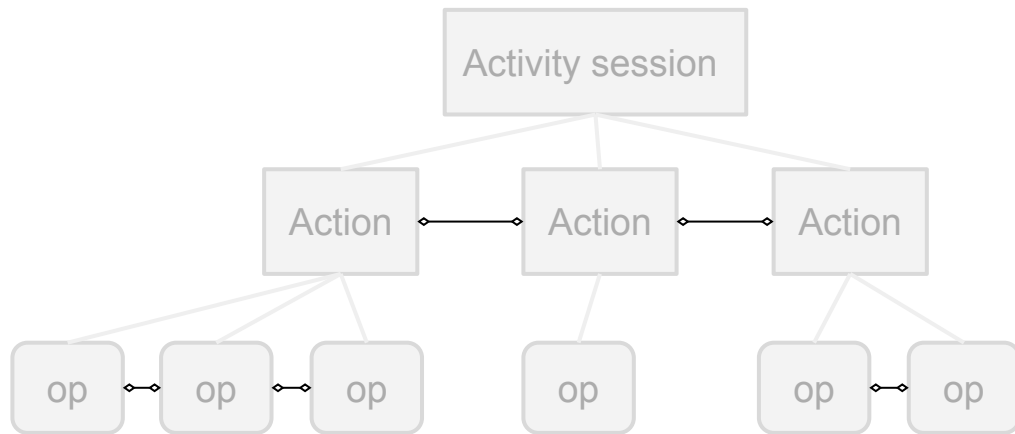
Clusters — short_within — within — between



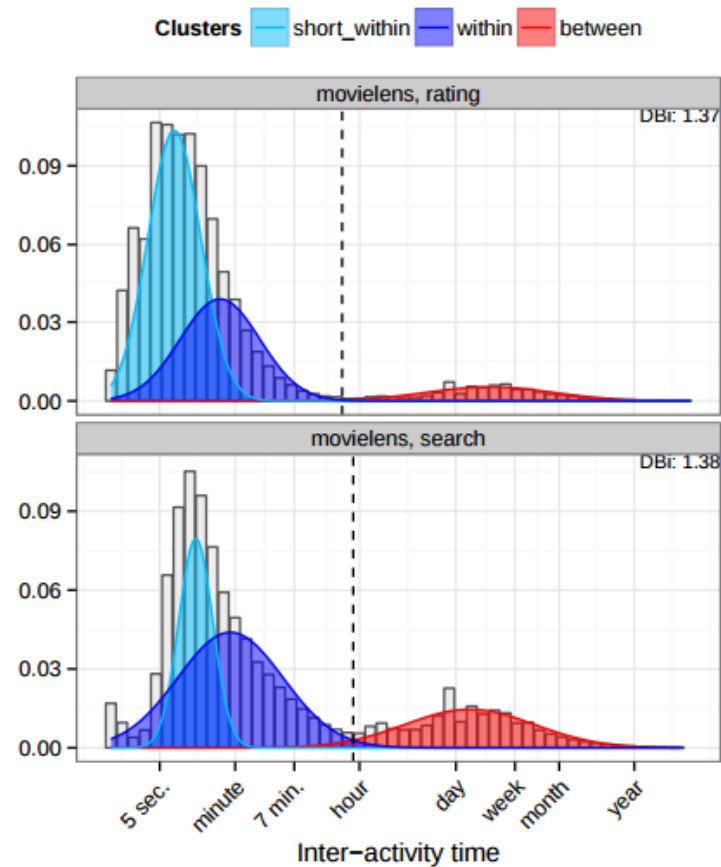


Short component

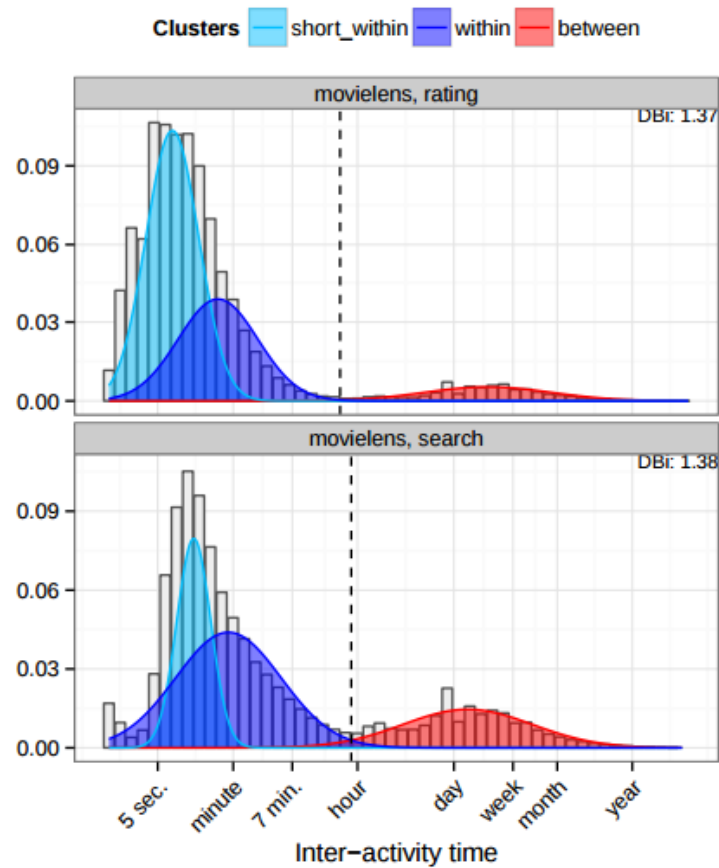
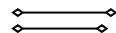




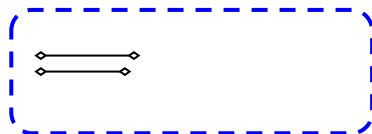
Short component



Short component



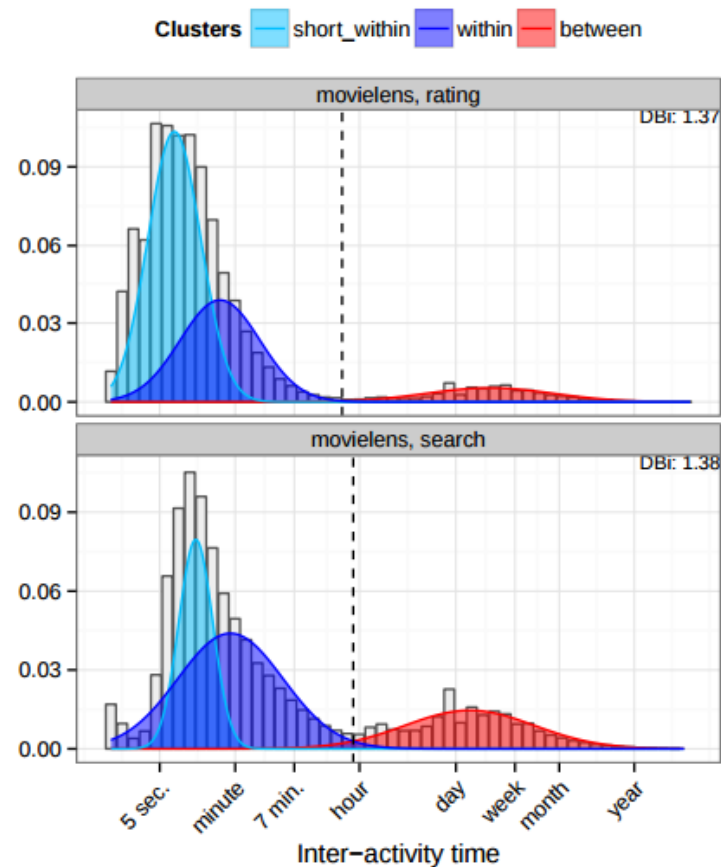
Short component



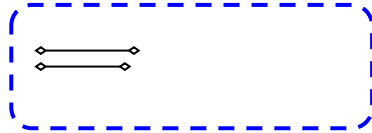
Action
(1-7 minutes)



Operation
(1-10 seconds)



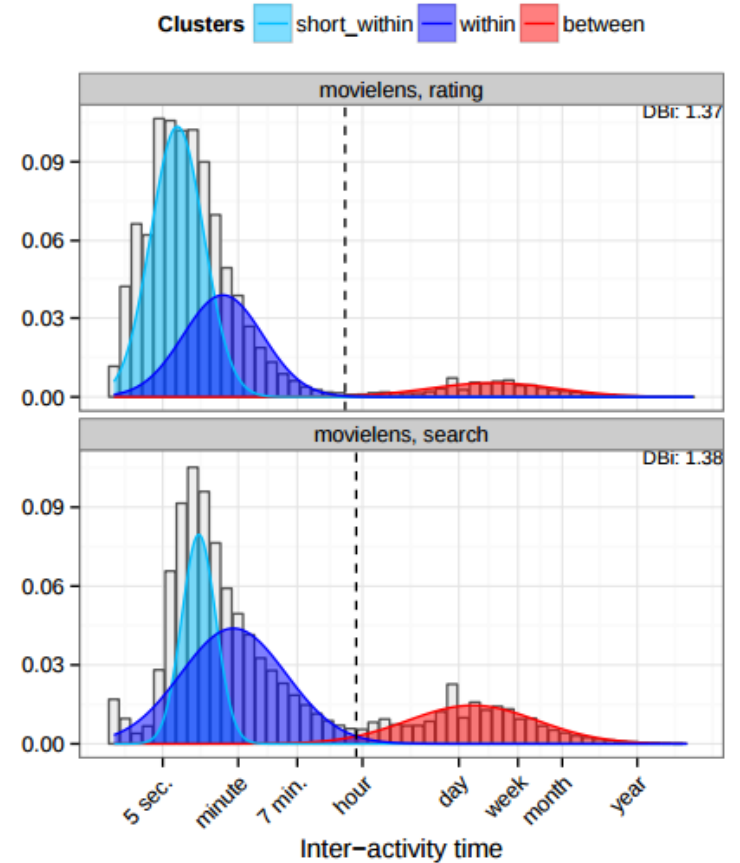
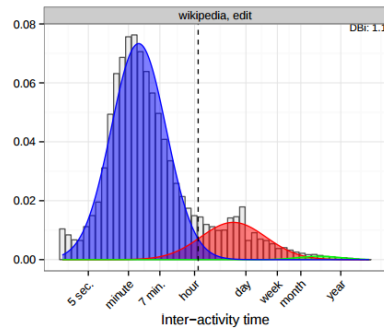
Short component



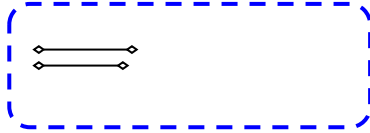
Action
(1-7 minutes)



Operation
(1-10 seconds)



Short component

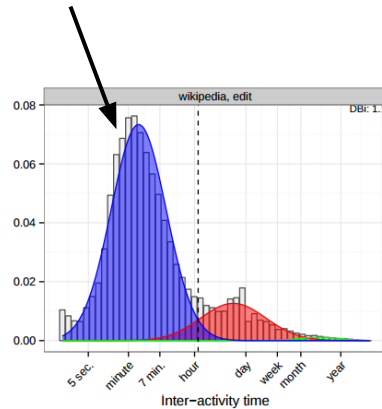


Action
(1-7 minutes)

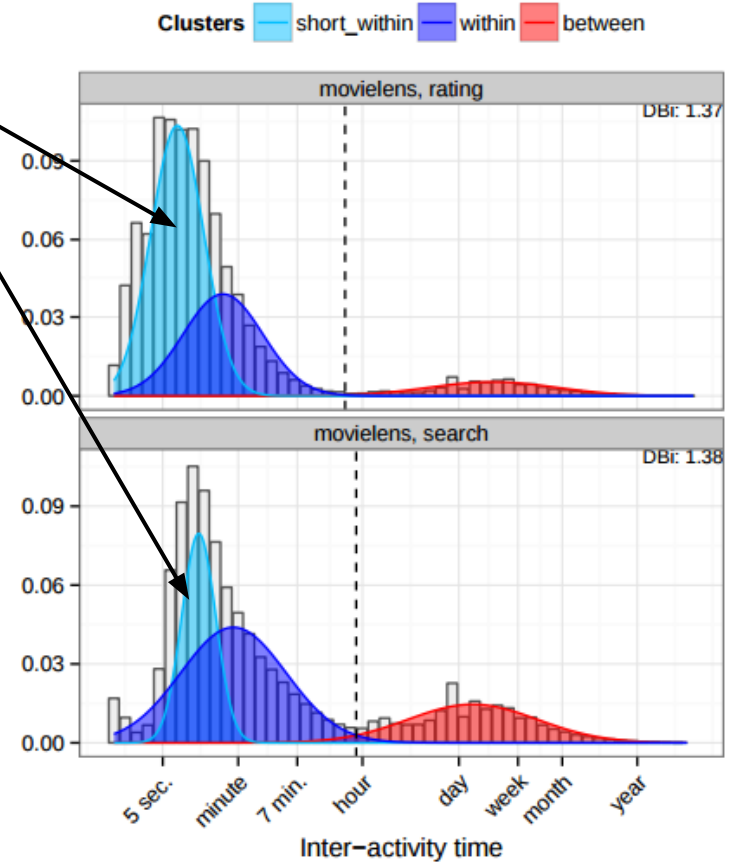


Operation
(1-10 seconds)

No AJAX



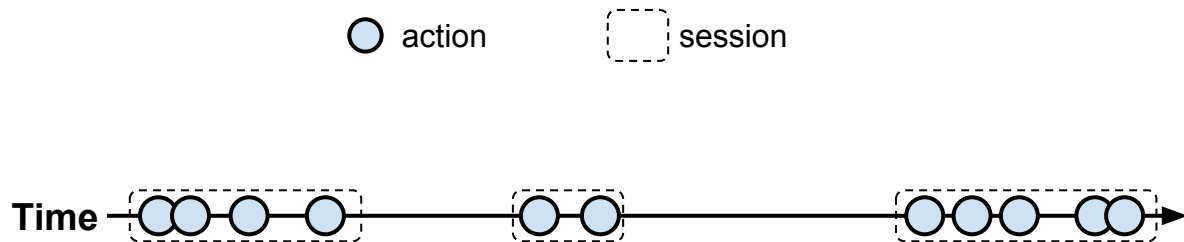
AJAX



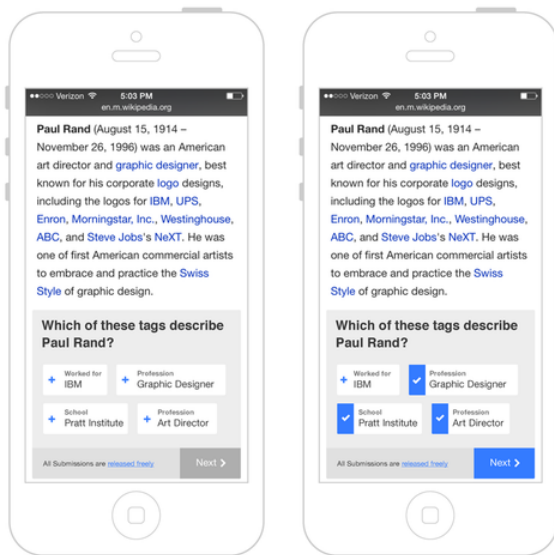
What's it all mean?

What's it all mean?

- sessions are *real things*
- 1 hour cutoff == good rule of thumb

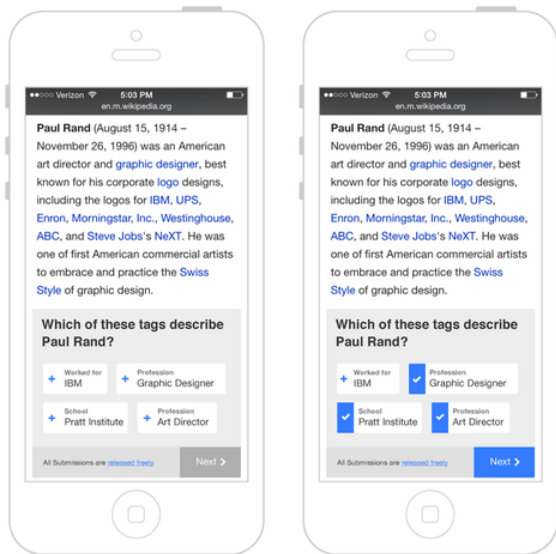


What's it all mean?

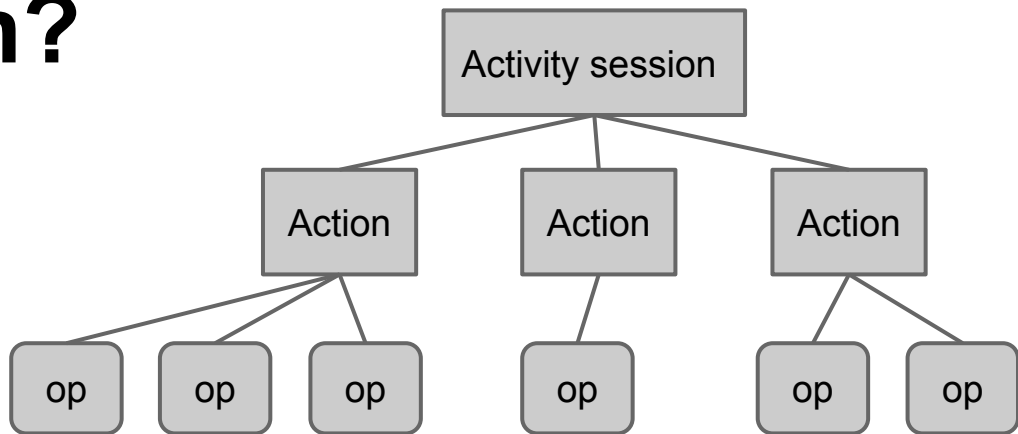


WikiGrok

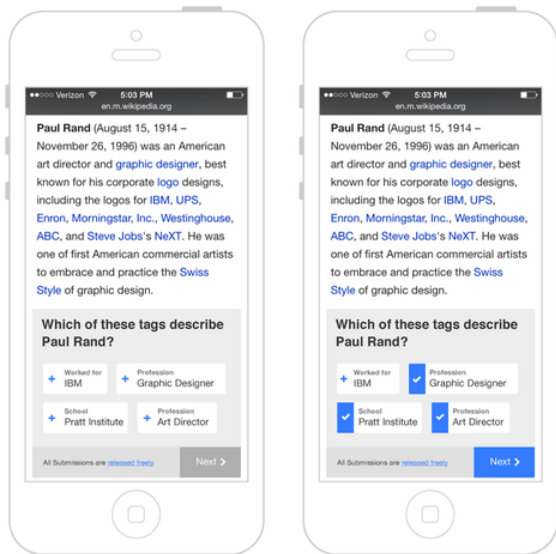
What's it all mean?



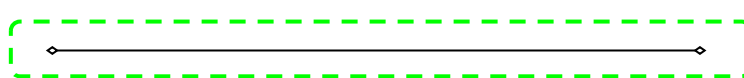
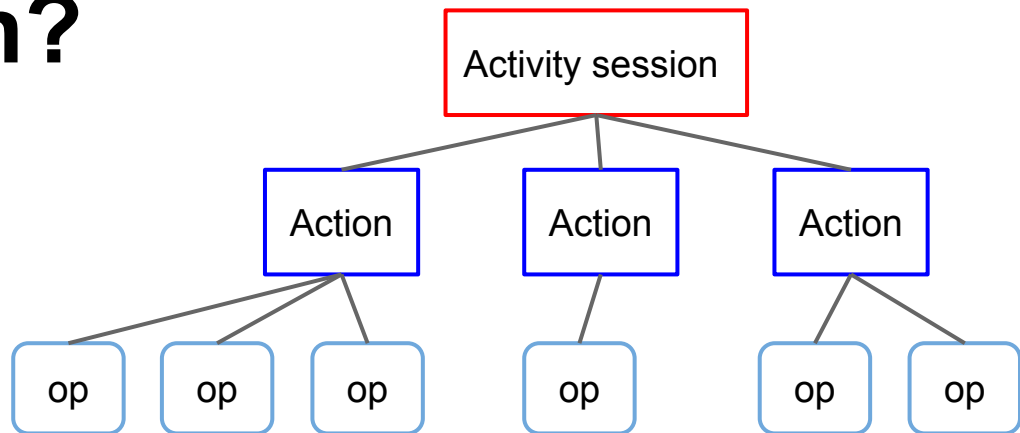
WikiGrok



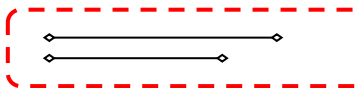
What's it all mean?



WikiGrok



Break
(1-12 months)



Between session
(12 hours - 1 week)



Action
(1-7 minutes)



Operation
(1-10 seconds)

Thanks

Aaron Halfaker

ahalfaker@wikimedia.org

twitter.com/halfak

User:EpochFail / User:Halfak (WMF)

“halfak” in #wikimedia-research

...to my collaborators:

Wikimedia Foundation

Oliver Keyes

GroupLens Research

Dan Kluver

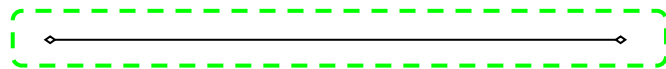
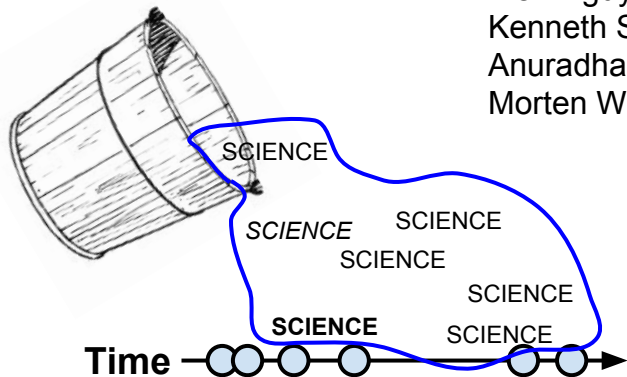
Jacob Thebault-Spieker

Tien Nguyen

Kenneth Shores

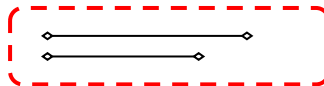
Anuradha Uduwage

Morten Warncke-Wang



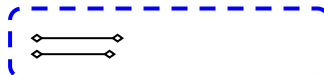
Break

(1-12 months)



Between session

(12 hours - 1 week)



Action

(1-7 minutes)



Operation

(1-10 seconds)

...and their NSF grants:

- 0808692
- 0964695
- 0968483
- 1017697

- 1111201
- 1218826
- 1319382